THE UNIVERSITY
of EDINBURGH

Quantitative models of biomolecular hydration thermodynamics

Georgios Gerogiokas

Doctor of Philosophy

University of Edinburgh

2015

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Georgios Gerogiokas*)

Abstract

This thesis explores the use of cell theory calculations to characterise hydration thermodynamics in small molecules (cations, ions, hydrophobic molecules), proteins and protein-ligand complexes. Cell theory uses the average energies, forces and torques of a water molecule measured in its molecular frame of reference to parameterise a harmonic potential. From this harmonic potential analytical expressions for entropies and enthalpies are derived. In order to spatially resolve these thermodynamic quantities grid points are used to store the forces, torques, and energies of nearby waters which giving rise to the new grid cell theory (GCT) model. GCT allows one to monitor hydration thermodynamics at heterogeneous environments such as that of a protein surface. Through an understanding of the hydration thermodynamics around the protein and particularly around binding sites, robust protein-ligand scoring functions are created to estimate and rank protein-ligand binding affinities. GCT was then able to retrospectively rationalise the structure activity relationships made during lead optimisation of various ligand-protein systems including Hsp90, FXa, scytalone dehydratase among others. As well as this it was also used to analyse water behaviour in various protein environments with a dataset of 17 proteins. The grid cell theory implementation provides a theoretical framework which can aid the iterative design of ligands during the drug discovery and lead optimisation processes, and can provide insight into the effect of protein environment to hydration thermodynamics in general.

Lay Summary

The work presented here is focused on the elucidation of hydration free energies in the context of protein and protein-drug systems. Essentially the thermodynamic stability of a water molecule in a particular space in a drug binding site is investigated. The thermodynamic stability is a measure of how much a water molecule prefers to be a in particular space when a system is at thermodynamic equilibrium. Understanding if a water molecule is stable in a particular binding site can then inform the drug discovery and optimisation processes. Knowing if a water molecule is very stable in a particular location can help understand how a drug candidate can be optimised. Knowledge of which pockets and subpockets of the binding site contains weakly bound waters can allow a medicinal chemist to design drugs which will kick out these waters. The stability of waters in particular spaces are calculated using grid cell theory (GCT). GCT uses the interatomic forces applied on a water molecule as well as interactions energies derived from a molecular dynamics simulations and decomposes the forces onto a grid to predict spatially resolved water thermodynamics. Essentially, the grid averages local water behaviour in a particular location in the binding site and gives an estimate of the thermodynamic stability near a grid point. Overall, the thesis first validates GCT by predicting the hydration free energy of ions, and other small molecules and comparing to experimental values. Afterwards, further studies were completed on several proteins, and protein-ligand systems focusing on understanding the nature of water thermodynamics in biomolecular systems with an emphasis on how this could aid drug discovery and optimisation.

Acknowledgements

It has been a long interesting journey! I would like to first thank Dr. Julien Michel for his guidance and insight into research. Dr. Richard Henchman was also very helpful throughout explaining cell theory and other insights. I would also like to thank Evotec UK Ltd. for funding this research. I would particularly would like to thank Dr. Michelle Southey for supervising me and everyone else at Evotec including Dr. Richard Law, Dr. Michael Mazanetz, Dr. Alexander Heifetz, Dr. Ewa Chudyk, Dr. Inaki Morao, Dr. Roger Robinson, Dr. Tim James and Dr. Michael Bodkin who were all very helpful, kind and inspiring people. They gave me insights into computer aided drug design and an introduction to methodologies used in their work. As well as this they gave very useful career advice during my many secondments in Oxfordshire which helped me gain a good all-around picture of the industry, techniques being utilised and the mindset required. I am also indebted to all members of the University of Edinburgh computational lab who have been good friends and a constant inspiration. In particular I would like to thank Gaetano Calabró for the advice and tolerating my endless questions on programming tips; Haris Georgiou for the many, many interesting conversations, and Kevin Pinto-Gil for the laughs. I also had many good conversations with everyone in the lab including Julien Sindt, Joshua Bradley-Shaw, Remi Cuchillo, Thomas Northey, Stefano Bosisio, Dr. Juan Bueren Calabuig and Pattama Wapeesittipan. I would also like to thank my old roommate, Dr. Michael Doig, who was always available for discussion and was a good buddy throughout the PhD. I would also like to acknowledge those outside the lab of course first the contribution of *Google* which answered any inane question with extreme tolerance. But on a more serious note, I am grateful to my parents for always supporting, raising and guiding me. Lastly, I am thankful to Maica Llavero who has always been a sun in my life.

Contents

Abbreviations

AM1-BCC	Semi-empirical (AM1) with bond charge correction (BCC)
AMBER	Assisted Model Building with Energy Refinement
APBS	Adaptive Poisson-Boltzmann Solver
CADD	Computer-aided drug design
CHARMM	Chemistry at HARvard Macromolecular Mechanics
EPW	Extra point for the additional shifted point charge found in TIP4P
FDTI	Finite difference thermodynamic integration
FMO	Fragment molecular orbital
GAFF	General atomic force field
GRID	Grid where different probe free energies are estimated
GROMACS	GROningen MAchine for Chemical Simulations
IFST	Inhomogenous fluid solvation theory
ITC	Isothermal calorimetry
JAWS	Just add waters
KS	Kolmogorov-Smirnov test
TIP	Transferable intermolecular potential functions
TIP3P	Transferable intermolecular potential function, 3 points
TIP4P	Transferable intermolecular potential function, 4 points
TIP4P-EW	TIP4P with Ewald correction
TIP5P	Transferable intermolecular potential function, 5 points
MC	Monte Carlo
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
NPT	Isothermal-isobaric ensemble
PBE	Poisson-Boltzmann equation
RESP	Restrained electrostatic potential
RISM	reference interaction site model
SASA	solvent accessible surface area
SPC	Simple point charge
SZMAP	Solvent-zap-map
VdW	Van der Waals

Symbols

DoF	Degree of freedoms
f	Force
G	Gaussian
i	Atom number
k_B	Boltzmann constant

m	Mass
N	Number of atoms
P	Pressure
$\mathbf{P}$	Probability
p	Momentum
r	Atomic position
T	Temperature
t	Time
δt	Time step
v	Velocity

Chapter 1

Introduction

"The map is not the territory" - Alfred Korzybski

1.1 Computer-aided drug design

Today, computational resources have become cheaper and faster allowing simulation of complex biological systems up to millions of atoms [1]. Such simulations produce trajectories which are the time evolution of the motions of the molecules resulting from Newtonian mechanics and molecular mechanic force fields. With these simulations molecular mechanisms can be investigated and the macroscopic properties such as temperature, pressure, and other thermodynamic parameters can be derived from the microstates of the system depending on the ensemble. Of key importance to the work presented here is the use of simulation methods for the dissection of molecular recognition in biological systems.

Protein-ligand molecular recognition is the non-covalent binding between ligands and protein targets. In medicinal chemistry the enthalpic aspects of the binding event are often described in terms of hyrogen-bonding (the interaction between two electronegative atoms with a hydrogen bonded to one of the atoms creating an attractive force), π interactions (interactions between π oribitals containing high electron density which can interact with other molecules or atoms) and hydrophobic association which are all represented by Lennard-Jones interactions and Coloumb interactions in a classical force field which are used for the work. In particular, molecular recognition between a protein and ligand is of interest in the pharmaceutical industry [2]. This is because activity modulation of a particular protein by ligand binding can create a therapeutic effect. Also in industry, ligand discovery and ligand optimisation can be expensive enterprises costing millions of pounds to deliver a drug to the market [3]. Computational modelling, if accurate, provides a method of lowering cost as well as developing an understanding and intuition of various ligand-protein systems [4].

1.1.1 Molecular recognition: a multi-objective optimisation problem

Understanding molecular recognition is key to the assessment of ligand hits in various drug discovery libraries. Hits are ranked as promising ligand binders found mostly in binding affinity assays. If ligands can be ranked and predicted *in silico* by their binding affinities then hit to lead optimisation and drug discovery processes can become more efficient [4].

Molecular recognition essentially involves finding a bioactive conformation of the biomolecular system of interest. It usually involves three major players; solvent, ligand and the protein (with the possible addition of ions and cofactors and rare DNA or RNA targets). In the 1960s until 1980s there was a dearth of computational resources which had led to conveniently assuming that the solvent, water, had a more neutral role in the whole process. This is why historically during the 1960s most progress in computationally-aided drug design (CADD) involved the assumption of shape and electrostatic complementarity which was reflected in interests in docking [5]. However, recently it has been shown the conformational entropy, small structural rearrangements, and solvent effects can greatly effect the molecular recognition event [6].

1.1.2 Ligand ranking problem - the key third player - water

One of the goals of CADD is to aid ligand optimisation/discovery by understanding where weakly bound and strongly bound waters are localised in a protein binding site. Weakly bound waters can be removed by new ligand modifications while tightly bound waters must be more carefully be considered for displacement [7, 8]. In the ligand-protein system there may be interacting waters which may or may not be removed through the modification of chemical moieties on the ligand. Also, water displacement is important for hydrophobic interactions which may occur as well as ligand and/or protein displacement of waters prior to polar interactions between the protein and ligand.

1.2 Water

Water is a very difficult liquid to study computationally because it is hard to model all of its properties. These are quite unusual compared to other simple liquids. It has unusually high freezing point and boiling point as well as surface tension and heat of vaporisation. all these properties may be linked with its strong polarity, polarisability and it ability to make hydrogen bonds. No single theoretical model is able to capture

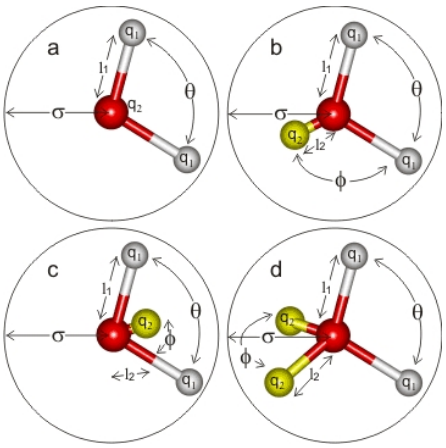


Figure 1.1: An overview of four of the main types of water models a) the 3 point model, e.g. SPC, TIP3P b) four point model with shifted point charge away from the hydrogen-oxygen-hydrogen angle, c) four point model with point charge moved closer to the hydrogen-oxygen-hydrogen angle, e.g. TIP4P, TIP4P-ew and d) a 5 point model, e.g. TIP5P. Image was adapted from the following source [12].

all of its thermodynamic and physical properties but usually can only capture a small range of its properties.

1.2.1 Anomalous properties of water

Many of the physical properties of water are anomalous compared to other liquids. This appears to be related to its unusual structuring. Some of these anomalous physical properties include expansion upon freezing, abnormally high heat capacity and high viscosity [9]. There are various theoretical water models which all are unable to completely capture all such properties. However, there have been new models which are being used to more accurately model a range of the properties of water with the addition of molecular simulation for appropriate sampling.

1.2.2 Water models

Many water models exist capturing some of the properties of water. Some of these involve the series of "TIP" (transferable intermolecular potential functions) waters [10] and SPC waters [11]. There are also more flexible and polarisable versions of water as well as coarse-grained waters.

Throughout the work here TIP4P-EW (transferable intermolecular potential functions 4 points with Ewald correction, figure 1.1c) was used exclusively. A recent study of water thermodynamics in bulk water suggests that TIP4P-EW is more accurate than the use of TIP3P (Figure 1.1a, three point model) [13] for density predictions. Another

more comprehensive study also agrees that TIP4P-EW predicts densities more closely to experimental reported densities [14]. Localisation of water is important to the work, so correct water bulk density would be a desirable starting point.

The TIP4P-EW model contains an extra EPW (extra point water) atom. This offsets the location of the charge on the oxygen and its parameters were optimised to correct for an Ewald Sum correction of the long range electrostatics and Lennard Jones interactions at various temperatures with associated experimental densities and enthalpies of vaporisation [13]. This model is more accurate than the simpler TIP3P and TIP4P models, and is shown to be better at estimating the oxygen-oxygen radial distribution function of bulk water [13]. It also estimates the free energy of vaporization, density and other properties quite well with appropriate sampling through Molecular Dynamics (MD) or Monte Carlo (MC) in bulk conditions [15].

However, of course there is need for work to compare the effects of different water models and see how predictions of molecular recognition event vary. For instance in a recent paper on a host-guest binding of cucurbituril-guest TIP3P gave better predictions of binding enthalpy in comparison to experimental data than with the TIP4P-EW water model used [16]. In another study hydration free energy calculations, free energy perturbation, were performed on 44 neutral small molecules using both TIP4P-EW and TIP3P water models [17]. This set of 44 small molecules included amides, amines, esters, alkanes, alkenes, alcohols, halides and thiols. This study shows that both TIP3P and TIP4P-EW are good water models for estimating hydration thermodynamics of neutral chemical groups. They both correlated well with experimental hydration free energies. For instance using AM1-BCC to parameterise partial charges of the small molecules and the TIP4P-EW and TIP3P water models for hydration thermodynamic calculations an R^2 of 0.93 and 0.94 respectively were found. However, water models also behave differently around charged moieties. In a work done by Joung and Cheatham [18], similar root mean square deviations are seen in the linear fit of predicted to experimental hydration free energies of halide and alkali ions between both TIP4P-EW and TIP3P models. So for the work done here TIP4P-EW was judged a good starting point because of its good water density; good correlation with experimental free energy of hydration on ions and neutral organic small molecules; and good free energy of vaporisation predictions.

1.2.3 Role of water in binding

Recently interest in solvent effects in protein-ligand systems has become more critical [19–22]. To analyse specific solvent effects molecular simulation is currently required because experimental techniques such as X-ray crystallography and NMR are usually unable to fully spatially resolve all relevant waters due to their mobility in the liquid

state and noisy signals in experiment. However, for very stable waters it is often possible to locate the oxygens of waters through either X-ray crystallography or NMR giving structural information on the less mobile waters. Molecular simulation can give insight into how water interacts in the protein-ligand system. Water is vital in the protein-ligand binding process whose thermodynamic contribution can be described by a thermodynamic cycle as show in Figure 1.2. The entire process is described by the following equation:

$$\Delta G_{\mathrm{b}} = -\Delta G_{\mathrm{hyd}}(P) - \Delta G_{\mathrm{hyd}}(L) + \Delta E_{\mathrm{int}} + \Delta G_{\mathrm{hyd}}(PL) \tag{1.1}$$

which contains the protein desolvation cost, ligand desolvation cost, interaction energy, and complex solvation energy but omits the conformational entropy of both the ligand and protein and the strain energy which may be induced in the ligand upon binding to the protein because it is computationally expensive to treat all conformational states.

Many of these costs are hard to obtain or inaccessible to experiment due to difficulties in dissecting the contributions which are often interdependent. For this reason the costs can usually only be obtained from simulation or theory. For example, there has been work using inhomogenous fluid solvation theory (IFST, further described later) which obtained estimates of the partial ligand and protein desolvation costs involved in the binding process. In the work by Breiten et al [23] the protein system human carbonic anhydrase was analysed on a congeneric (identical scaffold) ligand series. This was deemed a good test system because binding in this system results in little change in protein conformation (1 Å).

Protein and ligand partial desolvation costs; and complex solvation energies were found to be in the range of around -5 to 5 kcal mol^{-1} per water in this study. Interaction energies can cover a wide range of values depending on size they can typically range from 50-100 kcal mol^{-1} and usually make up a larger contribution. Finally rotational and translational entropy loss upon binding experimentally typically ranges from -5 to -22 kcal K^{-1} mol^{-1}. This sort of data is derived from molecular pair experiments where pairs are compared when covalently bound and separate [24]. Finally the the strain energy typically ranges from 0 to 39.7 kcal mol^{-1} from an analysis of crystal structures found of 33 compounds found in the PDB and Cambridge Structural Database [25]. All of these costs are important for understanding the binding process and the involvement of water. The justification for omitting the conformational entropy and strain energy was the fact that only congeneric ligands of the same scaffold where analysed where similar binding modes were adopted upon binding.

This is because water must be removed during the dewetting of the cavity of the protein ($\Delta G_{\mathrm{hyd}}(P)$) and the dewetting of the ligand ($\Delta G_{\mathrm{hyd}}(L)$) which must occur during the ligand binding process [26]. A general mechanism of dewetting has not been

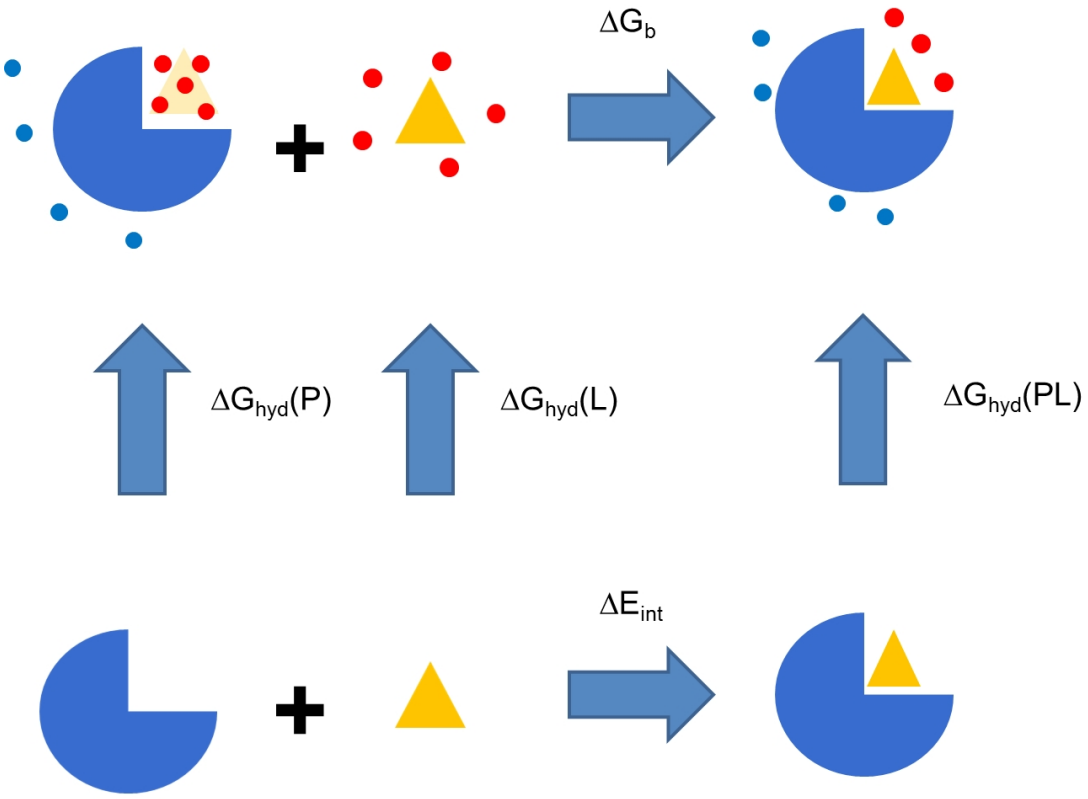


Figure 1.2: Thermodynamic cycle of a protein-ligand binding event where proteins are blue shapes and the ligand is a yellow triangle, in red circles are selected waters, and in blue circles are waters not included in the analysis. The protein desolvation cost ($-\Delta G_{\text{hyd}}(\text{P})$), ligand desolvation cost ($-\Delta G_{\text{hyd}}(\text{L})$), interaction energy ($\Delta E_{\text{int}}$) and protein-ligand solvation energy are included in the entire binding free energy, ΔG_{b}.

fully elucidated and is still being investigated but it seems likely the process is system dependent and can vary between different proteins [27]. Afterwards, there is the final hydration free energy in the bound state, $\Delta G_{\text{hyd}}(PL)$. The total contributions of all the desolvation contributions and the final solvation free energy of the bound state contribute to the total water reorganisation free energy.

One controversial issue is how hydration entropy is to be treated in such systems for each of these processes. The extent of the hydrophobic effect and entropy considerations has been hard to evaluate and is still a hotly contested debate [9]. This has to do with the lack of consensus on how to treat translational and rotational entropy of molecules of interest. There is both a system and molecular view on how to calculate the entropy of solutes and solvent. They treat the solute in different ways. The molecular view considers one molecule independently of the solvent and therefore requires a term to compensate for the effects of the surrounding solvent [22,28–30]. The system viewpoint does not distinguish between the solvent and solute by including a cratic entropy term which quantifies the number of minima of the solute in the solvent [31,32]. According to results by Irudayam et al. [31] the system approach tends to more closely match

the range of experimental entropy losses upon binding while the molecular view tends to overestimate these values. This seems to be because a large compensating entropy seems to be missing in a molecular approach such as IFST caused by an overestimate of entropy in solution [29] (which is later described in subsection 1.3.7). The inclusion of a solvent-exclusion term would increase the entropy putting the molecular approach more in line with experiment [31].

The overall water energetics play a vital role in biomolecular recognition, important for the hydrophobic association as well as often bridging interactions between a ligand and protein. Several studies on model cavities investigate water driven host-guest binding involving the dewetting and wetting processes during binding [33]. Dewetting is also investigated in other model systems such as the Cucurbit[7]uril system to develop an understanding of the role of water in protein-ligand recognition events [26, 34]. These studies are important for understanding how the thermodynamics of an association event.

There has also been investigations into the ranking of water sites in the protein or around the ligand. One study by Haider [35] investigates the role of water displacement in ligand optimization. This approach ranks water sites by the hydration free energy of each site using inhomogenous fluid solvation theory, (IFST), to give a displaceability score. Knowledge of favourable displaceable waters should aid ligand optimization during lead development. In order to gain this hydration thermodynamic behaviour in the binding site, molecular simulation is used. This can provide the spatial resolution necessary to identify water displacement costs in particular locations within the binding site. Such costs are necessary to assess the free energy cost of displacing a water where a new modification would interact with the protein. This water displacement cost must be overcompensated by the new protein-ligand interaction created (and new water interactions generated) in order for a stronger binder to be identified. However, highly thermodynamically stabilised waters would tend to be more difficult to displace and this cost must be compensated by new interactions the modified ligand would create in order to have an equivalent binder. So, in general strategies toward the targeting of less stable waters are used for improving binding. Spatially resolved thermodynamics provides a method for identifying easier methods to improve binding affinity of ligands inaccessible to experiment at the moment. Molecular simulation, and more specifically the sampling method of molecular dynamics is introduced as a method which can provide enough sampling to identify the free energies of waters of interest in a binding site.

1.3 Molecular simulation

1.3.1 Molecular dynamics

Molecular dynamics is the propagation of atomic positions of molecules using Newtonian dynamics according to the equation of motion

$$F_i(t) = m\ddot{r}_i(t) = -\frac{\partial U(r^N)}{\partial r_i} \tag{1.2}$$

This shows that the forces, F acting upon an atom i, is the negative of the partial derivative of the potential energy, U, with respect to the atomic position, r at time, t. Molecular dynamics is a function of the configuration of all the atoms and the potential energy of the system. With this methodology molecular and mechanistic details can be used to investigate particular collective behaviour (temperature, pressure, free energy, etc) as exhibited in the macroscopic scale. The potential energy of the system is defined by not only the positions of the atom but also their properties. These properties are defined by force fields.

1.3.2 Force fields

Ideally, a quantum mechanics (QM) representation where the electronic structure of molecular systems are taken into account would be used. However, this would restrict the study of larger biomolecular systems due to computational expense. For this reason, more computationally tractable molecular mechanics force fields are used throughout. These force fields have parameters which are derived from empirical data e.g. NMR $J-$coupling data for torsion angles, *ab initio* quantum calculations to approximate interaction potential terms for bonds, angles, dihedrals and electrostatics, as well as other macroscopic physical properties. The empirical force field then models the potential energy function of the system of interest.

Shown below is the potential energy of a general form of an AMBER molecular mechanics force field which was used throughout the work presented here.

$$U(r^N) = \sum_{\text{bonds}} \frac{k_i}{2}(l-l_0)^2 + \sum_{\text{angles}} \frac{k_i}{2}(\theta-\theta_0)^2 + \sum_{\text{torsions}} \frac{V_n}{2}(1+\cos(n\omega-\gamma)) + \sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\right) \tag{1.3}$$

The potential energy is a function of the configuration of the system which is defined by the atomic positions of all particles, r^N. In a typical biomolecular simulation the following terms are included in the force field using a simple harmonic potential, $(\frac{k_i}{2}(l-l_0)^2/\frac{k_i}{2}(\theta-\theta_0)^2)$ for the bond and angle terms. These oscillate around particular equilibrium lengths or angles found in experiment, l_0 and θ_0 respectively. The torsional term models how potential energy varies as a bond is rotated. This term often is parameterised by *ab initio* and NMR $J-$coupling data, V_n. The final term is composed of non-bonded interactions which involve van der Waal (vdW) interactions and electrostatics. Here ϵ_{ij} is the depth of the energy well, σ_{ij} is the distance where the potential between atoms i and j is zero, and r_{ij} is the distance. The vdW term is modelled by a Lennard-Jones potential which has an attractive component at longer distances $-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^6$ and a repulsive component at shorter distances, $\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}$. The Couloumbic term is simply given by the point charges q_i, q_j the distance between them, r_{ij} and the dielectric permittivity, ϵ_0 of the vacuum or solvent.

1.3.2.1 Particular force fields

Most protein force fields are fit to the previously discussed functional form. Many biomolecular force fields have been developed such as the CHARMM [36], GROMACS [37] and AMBER [38] force fields. They have been shown to replicate certain secondary structures and protein folding behaviours seen in experiment but sometimes exhibit biases to form certain types of secondary structures over others. For example, recent work by Beauchamp [39] has seen in a large dataset of proteins that the ff99sb-ildn-NMR force field has a larger propensity to form helices compared to experiment, while ff99sb-ildn-phi force field has a lower propensity than experiment. Both of these force fields are variants of the original ff99sb-ildn. The original ff99sb-ildn force field improved side-chain torsions of ff99sb by identifying which rotamers from alpha helical models deviate greatly from the torsions found in the PDB. The side-chain torsions were then fitted to new QM computed torsions [40]. The ff99sb-ildn-phi also adjusted the ϕ torsion (backbone) potentials of the ff99sb-ildn force fields which were optimised to take into account solvent interactions [41]. Finally the ff99sb-ildn-NMR used NMR backbone torsion measurements to optimise backbone torsions in combination with side-chain optimised torsions from ff99sb-ildn [42]. All of the force fields have their own particular

strengths and weaknesses reflecting how the protein force field was parameterised. In the majority of the work here the AMBER ff99sb [43] and ff12sb were used. These force fields are found to replicate reasonably well protein behaviour observed in numerous NMR experiments [39].

For ligands, the GAFF (general atomic force field) [44] was used. It has been thoroughly tested with AMBER force fields [44]. This is another AMBER force field which has generalised atom types for most organic compounds. The force field takes parameterised partial charges from semi-empirical methods such as AM1-BCC [semi-empirical (AM1) with bond charge correction (BCC)] [45] to model a particular ligand of interest. This force field was seen to perform best in combination with AM1-BCC in solvation free energy calculations for 241 ligands even compared to QM methods such as RESP (restrained electrostatic potential) used for the obtaining partial atomic charges [46].

1.3.3 Integrators

Integrators are required to propagate Newtonian dynamics during the molecular dynamic simulations to update velocities and positions for each particular time step. These integrators update the velocities and positions of atoms in the system. In the work here the velocity Verlet algorithm was used since it has a low integration error and is numerically stable [47]. Other variants such as simple Verlet and leapfrog Verlet methods work in similar ways with differences in how or when positions or velocities are updated.

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{f(t)}{2m}\delta t^2 \tag{1.4}$$

$$v(t+\delta t) = v(t) + \frac{f(t+\delta t) + f(t)}{2m}\delta t \tag{1.5}$$

Eq. 1.4 indicates the atomic position $r(t+\delta t)$ update step. Here the initial atomic position r at time t is updated in terms of the velocity $v(t)$ and contributions from the acceleration $\frac{f(t)}{m}$ to get the next atomic positions of all the atoms $r(t+\delta t)$. Afterwards the velocity for the next time step $v(t+\delta t)$ is updated for each atom in the system based on the potential energy, which gives the forces, and the masses of individual atoms. At this point one complete snapshot is generated. This update scheme is then reiterated during the entire MD simulation. Initial velocities are normally randomly assigned during the start of the MD simulation according to the Maxwell-Boltzmann distribution.

1.3.4 Thermostats

Experiments are typically carried out under conditions of constant atom number N, pressure P, and temperature T. For this reason it is necessary to have an algorithm to fix the temperature to any particular value desired, essentially a thermostat. The thermostats used here both rely on random numbers which are used to introduce noise into the system and maintain the average temperature, $T(t) = \sum_{i=1}^{N} \frac{m_i v_i^2(t)}{k_B \mathrm{Dof}}$ which is a function of the kinetic energy of atom i at time t, $\frac{1}{2} m_i v_i^2(t)$ and the number of degrees of freedom Dof which follows from the equipartition theorem, $\langle KE \rangle = \frac{1}{2} k_B T \times \mathrm{Dof}$.

1.3.4.1 The Andersen thermostat

The Andersen thermostat couples the system to a heat bath in order to maintain a particular temperature [48]. The heat bath is algorithmically just a series of stochastic forces acting randomly upon selected particles according to the particular Maxwell-Boltzmann distribution at a particular temperature.

$$\mathbf{P}(T; v) = v e^{-vT} \tag{1.6}$$

Equation 1.6, shows the probability $\mathbf{P}$, that a collision with the heat bath would occur for a temperature, T, which induces a new velocity, v. A simulation using the Andersen thermostat would first need to integrate the equations of motion. Secondly, a number of particles are selected according to the collision frequency and temperature. Finally, the particles will partake in a collision and a new velocity will be obtained from the Maxwell-Boltzmann distribution of the particular temperature.

1.3.4.2 The Langevin thermostat

Another commonly used thermostat is the Langevin thermostat [49]. This thermostat relies on the implementation of Langevin dynamics. Langevin dynamics relies on the change of momenta of particles in a system due to frictional interactions between the solute and solvent. Langevin equation is given by

$$m_i a_i(t) = -\xi v_i(t) + f_i^G \tag{1.7}$$

where the mass m, times acceleration a, given at time t is given from the frictionally damped velocity $-\xi v_i(t)$ and a random collision is partaken whose value follows a Gaussian G. Langevin dynamics this Gaussian can be used to to match a temperature. This is given in the following equation where

$$\mathbf{P}(\delta p_i^G) = \frac{1}{\sigma_p (2\pi)^{\frac{1}{2}}} e^{-\frac{1}{2}(\delta p_i^G/\sigma_p)^2} \tag{1.8}$$

the change in momenta is given by p. The change in momenta is a random variable which follows a Gaussian distribution G. This Gaussian is damped by the frictional coefficient ξ and has a probability that has the following form:

$$\sigma_p^2 = 2\xi m k_B T \int_t^{t+\delta t} f^2(t) dt \tag{1.9}$$

where m is the mass, k_B is the Boltzmann constant, T is the temperature and f is the force.

1.3.5 Barostats

Barostats maintain a selected pressure. This is necessary to sample the NPT statistical ensemble where the number of atoms, pressure and temperature are kept constant. This is necessary to compare with experimental data which typically are measured under NPT conditions. Ensembles are further discussed in section 1.4.2.

1.3.5.1 Isotropic Monte Carlo barostat

The isotropic Monte Carlo barostat adjusts the volume of the system by randomly scaling the box lengths. The volume move is then used to scale all the atomic coordinates of the system. The new distances between atoms results in a new energy of the system which is calculated. The box move is then accepted if the energy is accepted by a Metropolis criterion which matches the probability distribution of the statistical ensemble at the particular temperature chosen as shown in eqs 1.10, 1.11:

$$\Delta W = (U' - U) + P(V' - V) - N k_B T ln \frac{V'}{V} \tag{1.10}$$

$$\mathbf{P}(\Delta V) = \begin{cases} \mathrm{e}^{-\frac{\Delta W}{k_B T}}, \Delta W > 0 \\ 1, \ \Delta W \leq 0 \end{cases} \tag{1.11}$$

where U is the energy, P is the pressure, V is the volume, k_B is the Boltzmann constant, and N is the number of particles in the system.

1.3.5.2 Berendsen barostat

The Berendsen barostat obtains constant pressure by coupling to a temperature bath [50].

$$\mathbf{P}\left(\frac{\delta P}{\delta t}\right)_{\text{bath}} = \frac{P_0 - P}{\tau_p} \tag{1.12}$$

$$P = \frac{2}{3V}(E_k - \Xi) \tag{1.13}$$

Eq 1.12 and 1.13 shows the functional form of the barostat where Ξ is the virial from the pair-additive potentials (shown below, 1.14), P is the pressure, t is the temperature, F_{ij} is the interparticle forces, V is the volume, E_k is the kinetic energy, and τ_p is the time constant for the coupling.

$$\Xi = -\frac{1}{2}\sum_{i<j}\left\langle r_{ij} \cdot F_{ij} \right\rangle \tag{1.14}$$

The distances between molecules can be calculated using the centres of mass. Only the centres of mass of particles can be used consistently because intramolecular contributions to pressure are small in molecular systems [50]. The virial is the sum of pairwise forces weighted by their distance. The virial is changed to match the desired pressure by scaling interparticle distances, the r_{ij} component.

1.3.6 Electrostatic methods

There are two major methods of evaluating electrostatics in molecular simulations. These are the Ewald method and the reaction field method when using explicit waters. Here only the reaction field was used. There are also many implicit water models which treat waters as a continuum but these were not used because spatial resolution was of importance.

1.3.6.1 Reaction field

In the reaction field only atoms within a cutoff distance from an atom are explicitly treated. All areas beyond are treated as a dielectric continuum.

$$E = \frac{q_1 q_2}{4\pi\epsilon_0}\left(\frac{1}{r} + k_{rf} r^2 - c_{rf}\right) \tag{1.15}$$

$$k_{\mathrm{rf}} = \left(\frac{1}{r^3_{\mathrm{cutoff}}}\right)\left(\frac{\epsilon_{\mathrm{solvent}} - 1}{2\epsilon_{\mathrm{solvent}} + 1}\right) \tag{1.16}$$

$$c_{\mathrm{rf}} = \left(\frac{1}{r_{\mathrm{cutoff}}}\right)\left(\frac{3\epsilon_{\mathrm{solvent}}}{2\epsilon_{\mathrm{solvent}} + 1}\right) \tag{1.17}$$

Eqs. 1.15-1.17 shows how electrostatic energies are computed within the reaction field implementation where $\epsilon_{\mathrm{solvent}}$ is the dielectric permittivity constant of the solvent, r_{cutoff} is the spherical cutoff radii and q is the point charges of neighbours within the cutoff whose contributions are computed with the Coulomb law [51, 52]. In eq. 1.15 $\frac{1}{r}$ deals with the short range charges which are explicitly included within the cutoff and treated with the Coulomb Law, the k_{rf} term is the long-range term obtained from the dielectric area beyond and the last term is a correction term used to offset the interaction at zero at the reaction field cutoff. The c_{rf} term is an attempt to correct issues at the boundary between the reaction field and the dielectric area beyond. This correction works quite well as shown by the initial work by Tironi *et al.* [51]. They ran three simulations of 2127 SPC waters and 40 NaCl ions and compared the energy obtained. The energy of the electrostatics did not vary too much with a value of -140,800 (1000) at a reaction field cutoff of 9 Å compared to -138,200 (400) for the energy from the electrostatics with the root mean square fluctuations shown in the parentheses. There is a discontinuous jump in energy which leads to poor energy conservation when molecules move from outside to inside the reaction field. To avoid poor convergence tapering of interactions near the reaction field cutoff are implemented when $\epsilon_{\mathrm{solvent}} > 1$.

1.3.7 Alternative computational methods for the investigation of hydration energetics

Several molecular modelling techniques attempt to dissect the thermodynamic contribution of the solvent in protein such as IFST (inhomogenous fluid solvation theory) [26, 29, 30, 53], GRID [54], SZMAP (solvent-zap-map) [55], VISM (Variational Implicit-Solvent Model) [56, 57], RISM (reference interaction site model) [58], JAWS (Just add waters) [59], FMO (Fragment molecular orbital) [60] and cell theory [31, 61–64]

The solvent can be treated implicitly using models such as Poisson-Boltzmann, RISM and polarizable continuum models meaning the waters are not treated with an atomic

model but as a dielectric continuum. These can be used to estimate solvation free energies [65].

One of the earliest implicit solvation methods is called GRID [54]. In the 1980s, rational drug design was mainly concerned with the shape of the binding site when designing the ligand. GRID added energetic data for a water probe by creating energy contours around the binding site through the rolling of water probe along the surface accessible area of the protein. Calculation of Lennard-Jones terms, electrostatic and hydrogen bond terms were used to form an empirical energy function on a grid typically covering the binding site. Not only water probes but also methyl groups, amine nitrogens, carboxyl oxygens, and hydroxyl groups were used to probe different types of interactions that define the protein binding site. However, the approach did not consider entropic contributions to water stability.

Another implicit solvent model called SZMAP is discussed. The method adapts the Poisson-Boltzmann method and focuses on sites around the solute of interest. SZMAP treats the solvent as a dielectric continuum.

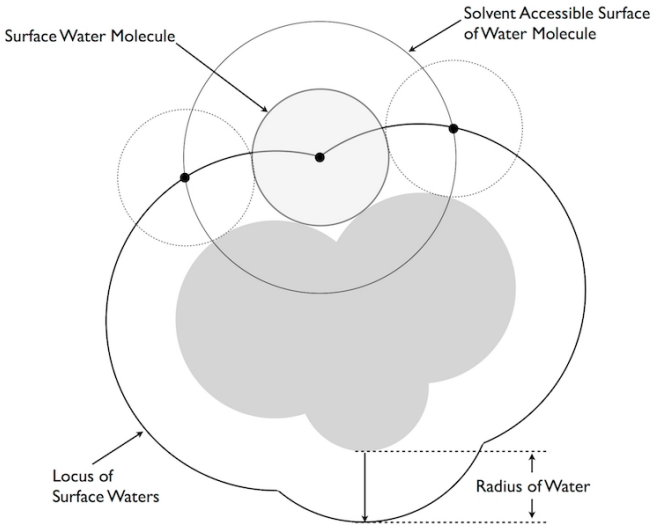


Figure 1.3: CUP8 Model showing how a spherical water probe is rolled around the molecule of interest, adapted from the Openeye website [66].

In this method a particular water model called CUP8 is used, see figure 1.3 [67]. The model is defined by a sphere which is an approximation of the surface area of a water molecule. The probe and protein has a dielectric constant of 1 (vacuum simulation) and a dipole moment of 1.86 debyes to match the properties of water in vacuum. In this model the cost of desolvation (free energy of transfer of a solute from vacuum to solvent) is estimated by having the probe water roll on solvent accessible surface area (SASA) of the solute of interest. Each accessible site is represented by a grid point at the center of the probe. At each of the grid points orientations of the probe are tested, with the directionality given by the dipole. Usually around 60 orientations are sampled

but 360 orientations are recommended for a more thorough sampling. It was later discovered that the curvature of the solute also affects the surface tension of the water and has macroscopic implications in water droplets. As the curvature increases, the surface area decreases and hydrophobicity increases because more energy is required to place the water at a specific site.

The benefit of SZMAP is that it can quickly and accurately define the free energy of a particular water binding site based on the electrostatic field. However, the components of the free energy are not completely separable into enthalpic and entropic components so it can be hard to rationalise the results. The method cannot detect water networks and water-water interactions which could be vital in many binding sites where water-water bridging is a common phenomenon.

VISM is an attempt to incorporate electrostatics from Poisson-Boltzmann theory as well as VdW surfaces of the protein and ligands, surface energies and an understanding of geometric effects into a clearer discrimination of hydration states in various chemical systems. It is essentially a mean-field free energy functional of the entire system described by its solvent-solute interfaces. To compute the free energy functional estimates of solute-solvent surface tensions at different interfaces are computed. The free energy functional contains solute-solvent interfacial energies, solute-solvent VdW energies, and electrostatic free energies. The end goal is to identify all free energy of hydration minima around protein environments [56]. The method has recently been tested on small molecule hydration thermodynamics as well as in a simple host-guest study involving cucurbit[7]uril and B2 ligand [57]. In both cases results qualitatively predict the experimental binding free energies. However, the method suffers from dependence on initial conditions. The hydration minima discovered depend on how an initial surface is generated.

As well as VISM, there has been another recent paper on 3D-RISM which now takes the cavity desolvation into account [58]. 3D-RISM produces a solvent distribution around a rigid solute. The method is also poor when dealing with hydrophobic hydration thermodynamics and for this reason a correction for cavity desolvation was introduced. 3D-RISM essentially is derived from the Ornstein-Zernike equation [58], which can allow spatial resolution of the solvent density distribution around a solute by evaluation of a susceptibility function for the solvent which uses specific atomic interaction potentials at different mixtures concentrations [68]. With the new correction term for the cavity 3D-RISM correlates hydration free energies of 504 small molecules yielding a R^2 value of 0.88 when compared to experimental data while free energy perturbation results yielded a R^2 value of 0.97. It is hoped that this will also be useful when looking at small molecules in binding site environments. As well as implicit solvent methods there are also QM methods being used to probe interesting parts of the binding site.

FMO is an *ab initio* method which assumes that the system can be broken down into fragments which, when added together will approximate the properties of the full system. The method assumes that particular chemical groups have a distinct local electron density [60]. Each fragment is treated separately in the electrostatic field of the remaining fragments of the system. When fragment energies are calculated the environmental Coulomb field is added by considering long-range interactions but the short ranged exchange and charge interactions are ignored. Afterwards, quantum-mechanical interactions are measured between pairs of fragments (dimers). Again the Coulomb field of the remaining fragments is retained. Any covalent bonds between fragments are fractionated without any capping because the Coulomb fields of the appending fragments are retained saturating these uncapped bonds.

With an FMO, water can be treated as clusters of two or one molecules and can probe the hydration free energy of particular areas of the binding site.

Furthermore, the interaction energies between fragments can be further decomposed using the PIEDA (pair interaction energy decomposition analysis) scheme [60].

$$\Delta E_{IJ}^{\text{int}} = \Delta E_{IJ}^{ES} + \Delta E_{IJ}^{\text{EXt}} + \Delta E_{IJ}^{\text{CT+mix}} + \Delta E_{IJ}^{\text{DI}} \tag{1.18}$$

In eq. 1.18 the interaction energy between a pair of molecules has been separated. There is the electrostatic (ES), exchange repulsion (EXt), charge transfer and higher order mixed terms (CT+mix) and dispersion (DI) terms. These terms can be used to measure the interaction energy of the water. They are derived from the polarisation terms from the monomer energies which gives a good estimate for the interaction energies. However, the method cannot easily consider entropic contributions because of computational expense. However, the energetic component is accurate at the QM level.

Many molecular simulation methods in combination with classical and quantum water models have been developed to predict hydration thermodynamics. One approach is JAWS (Just Add Water Molecules) [59]. This approach uses a Monte Carlo (MC) method, where the molecules are not simulated by Newtonian physics but undergo random moves which are then accepted or rejected according to a probabilistic criterion; typically the Boltzmann distribution is used. After successive moves the system can converge to the equilibrium distribution of the system. The JAWS method combines double decoupling theory and the lambda-dynamics method to dynamically insert and delete water molecules in a binding site of interest [69, 70]. The double decoupling method is described by eq 1.19 where $-\Delta G_{\text{hyd},(water)}$, the negative excess hydration free energy of a water molecule is added to $\Delta G_{\text{constr},(ideal,site)}$, the free energy of constraining an ideal gas to a site. Another term, the $\Delta G_{\text{trans},(water,site)}$, is used for the conversion of the localized ideal particle into a water molecule. The last

term $-\Delta G_{\mathrm{constr},(water,site)}$ is the free energy required to remove the constraints. This double decoupling method is then coupled to a lambda-dynamics method. In lambda-dynamics, a lambda parameter is used to alchemically transform ligand **1** into ligand **2** by using the parameter to alter the Hamiltonian based on ligand **1** to **2** or vice versa. Several intermediate lambda values can be used to facilitate the accuracy of the transition from one Hamilitonian to another. In this case a water model is shifted into a ideal water which cannot interact. In this way the binding free energy of waters in different regions of the binding site can be investigated allowing rational strategies for displacement of waters with weak binding free energies or lower entropies [71].

$$\Delta G_{\mathrm{b}(water)} = -\Delta G_{\mathrm{hyd},(water)} + \Delta G_{\mathrm{constr},(ideal,site)} + \Delta G_{\mathrm{trans},(water,site)} - \Delta G_{\mathrm{constr},(water,site)} \quad (1.19)$$

Another competing method for discovering hydration sites in proteins is IFST (inhomogenous fluid solvation theory). This method is based on work by Lazaridis [28–30] which established a method which views solvent changes from the perspective of the solute. The change in the solvent behaviour is then measured by molecular distribution functions. Molecular distribution functions are measures of how densities of molecules vary from a reference molecule. There can be pair correlation functions relating the change in densities between two molecules but there can also be the densities of three molecules and so on but these higher order molecular distribution functions are computationally expensive. From these molecular distribution functions correlations between atoms can then be described. The Kirkwood Buff solution theory can then be used to use these correlation functions (molecular distribution functions) to compute the thermodynamics of the liquid system. There have been further similar developments in the IFST field to what has been implemented in this work. One paper describes how the discretisation of the inhomogenous fluid solvation theory enables characterisation of water behaviours at defined volumes of the simulation space similar to the grid cell theory implementation. The new method is referred to as grid IFST (GIST) [26]. A discretisation of the space into k voxels (cubes) is defined usually in a cubic or rectangular grid of interest. In each of the voxels waters are allocated to particular grid points according to the location of nearest oxygen atoms. In each of these voxels Nguyen et al. defines the pair correlation function of waters to other waters and solutes it is interacting with locally. Then within the voxel the pair correlation function is treated uniformly because interpolation onto grid points currently uses a simple nearest point interpolation method. Each box will have its orientational and translational entropies defined from the local pair correlation function as well. The energies are then easily attainable from the configuration of the atoms in the particular box.

This method essentially discretises the simulation space and then truncates the molec-

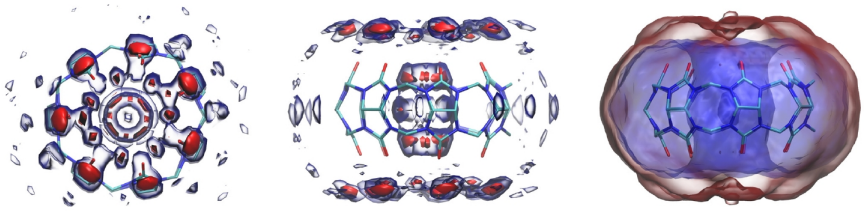


Figure 1.4: Grid IFST showing how waters prefer carbonyl oxygens in host cucurbit[7]uril model system (left and middle) where the free energy of hydration is shown with red being more stabilised and blue less stabilised. Right shows the entropy of the waters the central region are more ordered (bluer). Image from Nguyen et al. [72].

ular distribution functions at lower orders to obtain spatially resolved hydration thermodynamics. Figure 1.4 shows the free energy and entropy contours made after grid IFST was utilised.

1.4 Statistical mechanics of liquids

Statistical mechanics is introduced to explain the theoretical underpinnings of cell theory. Cell theory provides the foundations for GCT which was developed and used to obtain the spatially resolved hydration thermodynamics used throughout the work presented here.

The role of statistical mechanics is to connect microscopic states of a system to its macroscopic (aggregate) properties such as the temperature, volume and pressure of the system. One can consider a system of N particles. These particles have both momenta (p^N) and positions (r^N). These particles can be described in a $6N$ dimensional space, phase space, where the momenta of each atom have three components, p_x, p_y, p_z axes and coordinates positions along each x, y, z axis. Under particular conditions (such as fixed number of molecules, pressure and temperature), certain sets of positions and momenta will be adopted with varying probabilities which will influence the molecular properties. At equilibrium, the particles will have visited various microstates and will have a probability density $\mathbf{P}((p^N, r^N)_n)$ where n represents the number of sampled microstates. However, there is one major assumption required to be sure that one could sample phase space appropriately.

1.4.1 Ergodic hypothesis

The ergodic hypothesis assumes that all accessible microstates are sampled appropriately over the time evolution of the system at equilibrium. If this is the case a time average of a single molecular trajectory should sample all accessible microstates eventually. Without the ergodic hypothesis a relation between the time and ensemble

averages cannot be made. A good estimate can be achieved on a ns MD simulation but the amount of sampling required can be poor if there are large number of degrees of freedom and depending on how easy it is find rare events.

1.4.2 The *NPT* ensemble

In our simulations only the NPT ensemble is used because these are the conditions found in the lab which are to be compared to where the number of chemicals are constant, pressure is typically close to 1 atmosphere, and temperature is equal to 298 K. Much of the following text is adapted from ref [73].

From the NPT conditions certain ensemble properties can be defined. First, the Hamiltonian, H, is defined as the addition of the kinetic energy and potential energy of the system, which defines the total energy. From this it is commonly known that the probability density of the system is proportional to:

$$\mathbf{P}((p^N, r^N)_n) \propto V^N e^{-(H+PV)/k_B T} \tag{1.20}$$

This differs from the NVT (canonical ensemble) only by the additional PV term imposed by a constant pressure. The volume is now one of the microscopic quantities controlled by the pressure constraint during NPT conditions. From the probability density the partition function can be derived. Assuming Γ is a point in the phase space of the system:

$$Q_{NPT} = \sum_{\Gamma} \sum_{V} V^N e^{-(H+PV)/k_B T} \tag{1.21}$$

This partition function, Q_{NPT}, is a normalisation factor which sums all Boltzmann weighted microstates together. In the limit of infinite states this partition function can be rewritten as an integral:

$$Q_{NPT} = \frac{1}{N!} \frac{1}{h^{3N}} \frac{1}{V_0} \int dV \int d\mathbf{r} \int d\mathbf{p}\ e^{-(H+PV)/k_B T} \tag{1.22}$$

Here N is the number of atoms, h is Planck's constant and V_0 is the unit volume with an integration over all volumes, coordinates, and momenta.

With the partition function the free energy, or any other observable can be derived from the microstates. To get the free energy:

$$G = -\ k_B T \ln(Q_{NPT}) \tag{1.23}$$

In general from the sampled microstates the time averaged observable, $\langle A \rangle_{ens}$, can then be predicted assuming the system is ergodic and there is sufficient sampling of the states of phase space:

$$A_{obs} = \langle A \rangle_{ens} = \frac{1}{\tau_{obs}} \sum_{\tau=1}^{\tau_{obs}} A(\Gamma(\tau)) \tag{1.24}$$

1.5 Cell theory

Cell theory is a statistical mechanics theory initially created in the 1900s [74] and further developed throughout the years [75]. It was a theory which sprang from the experimental work on liquids investigated using X-ray diffraction [76]. Experimental data showed that in liquid argon there is short-ranged order similar to solids, which is maintained during the melting transition.

This led to an extension of lattice theory, a theory more associated with crystals. This was referred to as free volume or **cell theory** (used throughout). The theory assumes that a molecule, i, is confined in an occupied cell defined by its nearest neighbours. Within its cell each molecule moves in a field, $\psi(i)$, (and the mean-field potential, $\psi(0)$) which gives a partition function of:

$$Q_N = \lambda^{-3N} e^{-N\psi(0)/2k_B T v_f N} \tag{1.25}$$

where $\lambda = (h^2/2\pi m k_B T)^{\frac{1}{2}}$ otherwise known as the thermal de Broglie wavelength. With this the free volume, v_f can be calculated:

$$v_f = \int_{cell} e^{-(\psi(i)-\psi(0))/k_B T} dr_i \tag{1.26}$$

The supposition that the potential field can be defined by its nearest neighbours assumes a mean-field approximation (everything beyond the neighbours contribute equally). As well as this it is unnecessary to include correlations of motions beyond neighbours because it is often assumed in cell models that the molecule prefers to remain in the center of the cell where the field is exerted fully [77]. And finally some cell models such as those of Lennard-Jones and Devonshire [77] assume that the effective potential volume is defined by pair-wise potentials between the molecule within the cell and its neighbours around the sphere. However all these models do seem to have issues in the limit of low-density liquids:

$$\psi(r) \to 0 \text{ then } v_f \to \frac{V}{N} \text{ giving } Q_N = \lambda^{-3N} \left(\frac{V}{N}\right)^N \tag{1.27}$$

which suggests the addition of a Nk_B contribution to the entropy relating to a shared

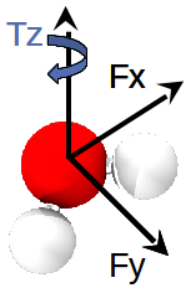


Figure 1.5: The analysis of forces and torques on water is portrayed along their respective x,y,z axes.

communal entropy which comes from the sharing of the cell volume. This leads to further developments including methods to directly add the communal entropy to the partition function. However, another suggestion resulted in hole theory where empty cells are included to account for the communal entropy.

Using the structure of the former cell theory (without any hole assumptions omitting the communal entropy since all cells are assumed to be occupied) [78] Henchman developed another cell theory for water which is not designed for low-density liquid studies. Henchman [31,61–64,78] approximates the potential energy surface of a water molecule using a six-dimensional anisotropic harmonic potential, shown in figure 1.5 to model three hindered rotations and translations. The harmonic potentials are parameterised by average force and torque constants derived from MD simulations. This differs greatly from the method of Lennard-Jones and Devonshire [77] which parameterises the model using nearest-neighbour potentials.

The theory was first validated and found to give results close to the experimental excess free energy of water using the TIP3P, TIP4P, TIP4P-EW, SPC, SPC/E and TIP5P models where results vary little (within 1 kcal mol^{-1}). Any errors in the enthalpic component are completely reliant on the accuracy of the force fields. However, the entropic components of the cell theory model have foundations on three major assumptions.

1. The partition function of the total system is the product of individual cell partition functions. Each individual cell partition function does not depend only on the coordinates within the cell. This is because in a simulation using force fields, all correlations are taken into account implicitly in the MD.

2. The six-dimensional anisotropic harmonic potential is assumed which only fails at low density and high temperatures due to greater translation and lower anisotropy.

3. Finally the force constant for each harmonic potential is obtained from the average ensemble average of the magnitude of the force, $\langle |f| \rangle$, which means that all correlations are implicitly taken into account. This is much more difficult in quasiharmonic analysis where a covariance matrix of the particle displacements of

a system has to have a defined reference frame, making implementation difficult, particularly for more complex systems [79].

The enthalpy of hydration is given here assuming an equivalent mole fraction solvation process neglecting the $P\Delta V$ which are typically negligible:

$$\Delta H_{X+w} = \left\langle U_{w(X)} \right\rangle - \left\langle U_{w(l)} \right\rangle - \left\langle U_{X(\mathrm{g})} \right\rangle \tag{1.28}$$

where U represents the ensemble average of the potential energies of a solvated solute, bulk water and the gas-phase solute respectively obtained from the MD simulations.

The enthalpy of hydration may also be written in terms of its intermolecular and intramolecular energies.

$$\Delta H_{X+w} = \left\langle U^{\mathrm{intra}}_{X,w(X)} \right\rangle - \left\langle U^{\mathrm{intra}}_{X(g)} \right\rangle + \frac{1}{2} \left\langle U^{\mathrm{inter}}_{X,w(X)} \right\rangle + \sum_{i=1}^{N_w} \left(\frac{1}{2} \left\langle U^{\mathrm{inter}}_{w,w(X)} \right\rangle - \left\langle U^{\mathrm{inter}}_{w(l)} \right\rangle \right) \tag{1.29}$$

The entropy of hydration is composed of six components.

$$\Delta S^{\circ}_{X+w} = \Delta S^{\circ,tr}_{X} + \Delta S^{\mathrm{rot}}_{X} + \Delta S^{\mathrm{int}}_{X} + \Delta S^{\mathrm{ori}}_{w,X} + \Delta S^{\mathrm{vib}}_{w,X} + \Delta S^{\mathrm{lib}}_{w,X} \tag{1.30}$$

In cell theory, hindered translations of the solute are accounted with a three-dimensional harmonic potential, and the solute translational entropy given by eq. (1.31).

$$\Delta S^{\circ,\mathrm{tr}}_{X} = k_B \ln \left\{ \frac{1}{V_w} \prod_{j=1}^{3} \frac{2 k_B T e^{1/2}}{\left\langle F^{j}_{X(\mathrm{aq})} \right\rangle} \right\} \tag{1.31}$$

where k_B is the Boltzmann constant, V_w is the volume available to water in the gas state, e is the base of the natural logarithm, and j is a principal axis in the solute frame of reference. $\left\langle F^{j}_{X(\mathrm{aq})} \right\rangle$ is the ensemble average of half the magnitude of the forces along each principal axis. Half the magnitude is used to avoid double counting and to bring the from a frame of reference of a molecule into that of the system.

The rotational entropy has a similar form:

$$\Delta S^{rot}_{X} = min \left(0, k_B \ln \left\{ \frac{1}{V_w} \prod_{j=1}^{3} r^{j}_{X} \frac{2 k_B T e^{1/2}}{\left\langle \tau^{j}_{X(aq)} \right\rangle} \right\} \right) \tag{1.32}$$

where $\left\langle \tau^{j}_{X(aq)} \right\rangle$ is the ensemble average of half the magnitude of the torques along

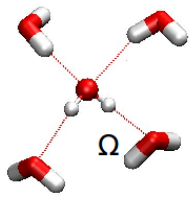


Figure 1.6: Orientational entropy is computed by a coordination number which is defined are all water molecules within a cutoff of 3.4 Å (matching the typical range of the first solvation shell of a water). Here a cluster of waters are shown with neighbours hydrogen bonding a central water.

each principal axis, and r_X^j is the distance from the principal axis to the vdW surface. The components inside the natural logarithm account for the orientational minima of the particular solute. The min() function is used to avoid cases where the harmonic approximation breaks down. This typically happens with small weakly interacting solutes where orientational minima overlap would otherwise lead to an increase rotational entropy upon hydration, a nonphysical result. For this reason the minimum entropy change is assumed to be 0 kcal mol^{-1} K^{-1} meaning $\Delta S \geq 0$.

The internal entropy of the solute is computed using eq. 1.33. This is essentially the ratio of the Gibbs entropy of the internal coordinates of a solute conformations in the solution and those in the gas phase.

$$\Delta S_X^{\text{int}} = -k_B \int \frac{\rho(\text{r})_{w(X)} \ln \rho(\text{r})_{w(X)} d\text{r}}{\rho(\text{r})_{X(gl)} \ln \rho(\text{r})_{X(g)} d\text{r}} \tag{1.33}$$

The orientational entropy of N_w waters is defined as:

$$\Delta S_{w,X}^{\text{ori}} = N_w k_B \ln \left\{ \frac{\left\langle \Omega_{w(X)}^{\text{ori}} \right\rangle}{\left\langle \Omega_{w(l)}^{\text{ori}} \right\rangle} \right\} \tag{1.34}$$

where $\left\langle \Omega_{w(l)}^{ori} \right\rangle$ is the ensemble average of water orientations in bulk which varies according to water model and simulation conditions while $\left\langle \Omega_{w(X)}^{ori} \right\rangle$ is the ensemble average of water orientations near a solute given by eq. (1.35).

$$\left\langle \Omega_{w(X)}^{\text{ori}} \right\rangle = \left\langle \frac{1}{N_w} \sum_{i=1}^{N_w} \Omega_{w,i}^{\text{ori}} \right\rangle \tag{1.35}$$

To approximate the orientational entropy a generalised Pauling model is used as shown in equation (1.36), and (1.37), and as illustrated in figure 1.6. Eq. (1.36) is used unless a water is in a coordination shell of a polar solute atom.

$$\Omega_{w,i}^{\text{ori}} = \frac{N_{a,i}(N_{a,i}-1)}{2}\left(\frac{N_{a,i}-2}{N_{a,i}}\right)^2 \tag{1.36}$$

$$\Omega_{w,i}^{\text{ori}} = \frac{N_{a,i}^{\text{eff}}(N_{a,i}^{\text{eff}}-1)}{2}\left(\frac{N_{w}^{\text{bulk}}-2}{N_{w}^{\text{bulk}}}\right)^{2-p_{\text{HB}}^{x}} \tag{1.37}$$

$N_{a,i}$ is the number of hydrogen bond acceptors (a) about the coordination shell of water i of which here a cutoff of 3.4 Å is used. N_w^{bulk} is the coordination number of water in bulk. $N_{a,i}^{\text{eff}}$ is effective coordination number which can account for solute hydrogen bond acceptors as well in the local environment. This is done by taking various contributions to the $N_{a,i}$, number of acceptors into the acceptors from the solute, $N_{X,i}$, acceptors from the first shell waters, $N_{w_s,i}$, and other waters beyond the coordination shell of the solute, $N_{w_b,i}$.

$$N_{a,i} = N_{X,i} + N_{w_s,i} + N_{w_b,i} \tag{1.38}$$

Next the probability of each type of acceptor being hydrogen bonded to water i is then defined, in eq. 1.39. Finally, these probabilities can be combined to compute the effective coordination number, eq. 1.40.

$$p_{\text{HB}}^{X} = \frac{N_{X\text{HB},i}}{N_{X,i}}; p_{\text{HB}}^{w_s} = \frac{N_{w_s\text{HB},i}}{N_{w_s,i}}; p_{\text{HB}}^{w_b} = \frac{N_{w_b\text{HB},i}}{N_{w_b,i}} \tag{1.39}$$

In this equation hydrogen bonds are defined simply by a force definition in which the nearest donor to an acceptor is deemed to be the donor where the largest value of q_a/r_{AH}^2 is the largest where q_a is the partial atomic charge of the atom A and r_{AH} is the distance between the donor and the acceptor.

$$N_{a,i}^{\text{eff}} = \frac{N_{X\text{HB},i} + N_{w_s\text{HB},i} + N_{w_b\text{HB},i}}{\max(p_{\text{HB}}^{X}, p_{\text{HB}}^{w_s}, p^{w_b}_{\text{HB}})} \tag{1.40}$$

However, this formulation of cell theory is not able to spatially resolve the thermodynamics in particular areas of simulation space. In this work grid cell theory (GCT) is proposed as a novel statistical mechanics methodology for interpreting the effect of the solvent in an intuitive and visual manner.

GCT discretises simulation space onto a grid which contains the entire protein-ligand system. This allows decomposition of entropy/enthalpy of the free energy of the ligand-binding process. Also, it is of key importance in predicting and understanding the effect

of buried waters on the binding modes of ligands and the free energy of binding which is further described in the next chapter.

Chapter 2

Grid Cell Theory: discretisation of cell theory

"The library is a sphere whose exact center is any one of its hexagons and whose circumference is inaccessible - Jorge Luis Borges"

2.1 Theory

Grid cell theory discretises cell theory into a spatially resolved grid which can be uniform or uneven if desired. Here a cubic evenly spaced grid is positioned around the solute X in a volume of space $\mathbf{s}$, where the grid is located. To speed up convergence only one conformation r for each small molecule is restrained with harmonic positional restraints on all heavy atoms (but they may even be fixed). However, if solute flexibility is required multiple conformations can be restrained in a similar manner. These restraints help speed up convergence and allow clearer visualisation of the grid outputs.

Because the ligands are rigid, intramolecular energies are cancelled and the contribution to the enthalpy in $\mathbf{s}$ is given by eq. (2.1):

$$\Delta H^{\mathbf{s}}_{X(\mathbf{r})+w} = \left\langle \frac{1}{2} U^{\text{inter}}_{X(\mathbf{r}),w(X(\mathbf{r}))} \right\rangle + \left\langle \sum_{i=1}^{N^{\mathbf{s}}_w} \left(\frac{1}{2} U^{\text{inter}}_{w_i,w(X(\mathbf{r}))} - U^{\text{inter}}_{w(l)} \right) \right\rangle = \Delta H^{\mathbf{s}}_{X(\mathbf{r})} + \Delta H^{\mathbf{s}}_{w} \tag{2.1}$$

$N^{\mathbf{s}}_w$ is the number of the waters in the volume $\mathbf{s}$. The intermolecular interactions of atoms within and outside of area $\mathbf{s}$ is taken into account. The first term, is the ensemble average of the intermolecular energies between the first shell waters, $w(X(\mathbf{r}))$ and the solute, $X(r)$, in its conformation $\mathbf{r}$ divided by two to avoid double counting. The second term is the ensemble average of waters, w, in the space, $\mathbf{s}$ between first shell waters

and further. Each of these waters are subtracted by $U^{\text{inter}}_{w(l)}$ the average intermolecular of bulk water obtained from pure bulk water simulations. Next we have the equations for solute entropy. Similarly to the solute intramolecular energies, the internal solute entropy cancels because of the rigid solute assumption, and so is not shown. The solute entropy is composed of the translational and rotational entropies as shown in eq. 2.2.

$$\Delta S^{\text{s},\circ}_{X(\mathbf{r})+w} = \Delta S^{\text{s},\circ,\text{tr}}_{X(\mathbf{r})} + \Delta S^{\text{s,rot}}_{X(\mathbf{r})} + \Delta S^{\mathbf{s},w}_{w} \tag{2.2}$$

$$\Delta S^{\text{s},w}_{w} = \Delta S^{\text{s,ori}}_{w} + \Delta S^{\text{s,vib}}_{w} + \Delta S^{\text{s,lib}}_{w} \tag{2.3}$$

The $\mathbf{s}$ volume is further discretised into N_s voxels, k, of volume *V(k)* for which each contains cell parameters computed as ensemble averages for eqs. 2.4-2.9.

$$\Delta H^{\text{s}}_{X(\mathbf{r})}(k) = \left\langle \frac{\sum_{i=1}^{N^{\text{s}}} \frac{1}{2} U^{\text{inter}}_{w_i,X(r)} I(k)}{\max(1, \sum_{i=1}^{N^{\text{s}}}) I(k)} \right\rangle \tag{2.4}$$

$$\Delta H^{\text{s}}_{w}(k) = \left\langle \frac{\sum_{i=1}^{N^{\text{s}}_w} \left(\frac{1}{2} U^{\text{inter}}_{w_i,w(X(r))} - U^{\text{inter}}_{w(l)} \right) I(k)}{\max(1, \sum_{i=1}^{N^{\text{s}}_w} I(k))} \right\rangle \tag{2.5}$$

$$\Delta \Omega^{\text{ori}}_{w}(k) = \left\langle \frac{\sum_{i=1}^{N^{\text{s}}_w} \Omega^{\text{ori}}_{w,i} I(k)}{\max(1, \sum_{i=1}^{N^{\text{s}}_w} I(k))} \right\rangle \tag{2.6}$$

$$F^{j}_{w}(k) = \left\langle \frac{\sum_{i=1}^{N^{\text{s}}_w} F^{j}_{w,i} I(k)}{\max(1, \sum_{i=1}^{N^{\text{s}}_w} I(k))} \right\rangle \tag{2.7}$$

$$\tau^{j}_{w}(k) = \left\langle \frac{\sum_{i=1}^{N^{\text{s}}_w} \tau^{j}_{w,i} I(k)}{\max(1, \sum_{i=1}^{N^{\text{s}}_w} I(k))} \right\rangle \tag{2.8}$$

$$\rho(k) = \left\langle \sum_{N^{\text{s}}_w}^{i=1} \frac{I(k)}{V(k)\rho_b} \right\rangle \tag{2.9}$$

For all these parameters *I(k)* is an indicator function which is 1 if the oxygen atom of a water molecule is within the voxel otherwise it is 0. $\rho(k)$ is the relative water density compared to bulk ρ_b which depends on the particular water model and simulation conditions. The denominator of each equation is used to normalise the cell parameters to have per-water statistics.

The eqs. 2.10-2.12 define the number of waters per voxel, k, and space $\mathbf{s}$, and lastly the relative density of $\mathbf{s}$.

$$N_w(k) = \rho(k)V(k) \tag{2.10}$$

$$N_w(\mathbf{s}) = \sum_{k=1}^{N_\mathrm{s}} N_w(k) \tag{2.11}$$

$$\rho(\mathbf{s}) = \sum_{k=1}^{N_\mathrm{s}} \rho(k)/N_\mathrm{s} \tag{2.12}$$

Furthermore there are the eqs. 2.13-2.16 for the solute and solvent hydration enthalpies of $\mathbf{s}$. These values can be normalised as per water values.

$$\Delta H^\mathrm{s}_{X(\mathbf{r})} = \sum_{k=1}^{N_\mathrm{s}} N_w(k)\Delta H^{(\mathrm{s})}_{X(r)}(k) \tag{2.13}$$

$$\Delta H^{n,\mathrm{s}}_{X(r)} = \Delta H^\mathrm{s}_{X(r)}/N_w(\mathrm{s}) \tag{2.14}$$

$$\Delta H^\mathrm{s}_w = \sum_{k=1}^{N_\mathrm{s}} N_w(k)\Delta H^{(\mathrm{s})}_w(k) \tag{2.15}$$

$$\Delta H^{n,\mathrm{s}}_w = \Delta H^\mathrm{s}_w/N_w(\mathrm{s}) \tag{2.16}$$

One can get the average orientational number, forces, and torques, from eqs. 2.17-2.19 respectively:

$$\Omega^\mathrm{ori}_w(\mathrm{s}) = \max\left(1, \frac{1}{N_w(\mathrm{s})}\sum_{k=1}^{N_\mathrm{s}} N_w(k)\Omega^\mathrm{ori}_w(k)\right) \tag{2.17}$$

$$F^j_w(\mathrm{s}) = \frac{1}{N_w(\mathrm{s})}\sum_{k=1}^{N_\mathrm{s}} N_w(k)F^j_w(k) \tag{2.18}$$

$$\tau^j_w(\mathrm{s}) = \frac{1}{N_w(\mathrm{s})}\sum_{k=1}^{N_\mathrm{s}} N_w(k)\tau^j_w(k) \tag{2.19}$$

The orientational number cannot be lower than one, which would be found in the ideal gas state. With these cell parameters the orientational, librational and vibrational entropies can be computed for **s**, from eqs. 2.20-2.22, which can also be normalised to per-water values.

$$\Delta S^{\mathrm{s,ori}}_{w,X(\mathbf{r})} = N_w(\mathrm{s}) k_B \ln \left\{ \frac{\Omega^{\mathrm{ori}}_w(\mathrm{s})}{\Omega^{\mathrm{ori}}_{w(l)}} \right\} \tag{2.20}$$

$$\Delta S^{\mathrm{s,vib}}_{w,X(r)} = N_w(\mathrm{s}) k_B \ln \left\{ \prod_{j=1}^{3} \frac{F^j_{w(l)}}{F^j_w(\mathrm{s})} \right\} \tag{2.21}$$

$$\Delta S^{\mathrm{s,lib}}_{w,X(r)} = N_w(\mathrm{s}) k_B \ln \left\{ \prod_{j=1}^{3} \frac{\tau^j_{w(l)}}{\tau^j_w(\mathrm{s})} \right\} \tag{2.22}$$

Finally, the contribution of the waters in **s** can be computed to yield the free energy change of hydration of the solute X, given by eq. 2.23

$$\Delta G^{\mathrm{s}}_{w,X(r)} = \Delta H^{\mathrm{s}}_{X(r)+w} - T\Delta S^{\mathrm{s},w}_{w} \tag{2.23}$$

This theory was first validated with bulk water followed by a dataset of small molecules of various charges and polarities. A dataset of small molecules was analysed and showed good correspondence with experimental hydration free energies. This is fully described in chapter 3. This was to confirm the validity of its use in biomolecular systems which would contain a broad range of chemical environments, including polar, charged, and nonpolar.

2.2 Nautilus workflow

Grid cell theory has been implemented into a python code called *Nautilus*. A typical workflow for a GCT computation is described in appendix A. Each computation involves the post-processing of a simulation of either a solute, ligand, protein or a complex. Often the protein configuration found when the ligand is bound rather than a relaxed protein configuration in the solvent is used as a reference. Typically these structures are derived from a X-ray crystallography or NMR structures. In the simulations either a TIP4P-EW or TIP3P water model is supported but only TIP4P-EW has been tested extensively. Any protein force field may be accommodated but only combinations of TIP4P-EW with ff99SB and TIP4P-EW with ff12SB have been tested within the work

here. At the end of the calculation there are grids for hydration free energy, enthalpy, entropy, and their components as well as relative water density.

2.3 Choosing regions of interest

Once a grid has been generated there are several ways of selecting regions.

2.3.1 Selecting from solute centre

With the solute center method one can simply select grid points as a distance from the centre. Cubic or spherical regions are taken about the centre of the solute atom which is typically restrained in the simulation with a harmonic positional restraint (force constant, 10 kcal $mol^{-1}Å^{-2}$).

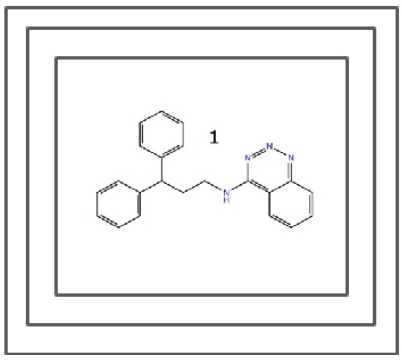


Figure 2.1: Cubic method defines regions of grid points as function of distance from any coordinate specified, typically one at the centre of mass of the solute of interest

From the coordinate of the restrained atom grid points are selected either with a cubic or spherical distance cutoff (an example of a cubic cutoff is shown in figure 2.1). As the spherical or cubic region gets larger conditions at the edges become more bulk-like which contributes negligibly to the hydration free energy and leads to a converged value. In practice, if larger volumes are monitored this introduces greater fluctuations which cause convergence to slow down due to the larger numbers of waters. Spherical regions tend to converge faster than cubic regions because a smaller volume is monitored. This is further discussed in the next chapter.

As well as simple restrained coordinates of a solute, clustered grid points can also be used.

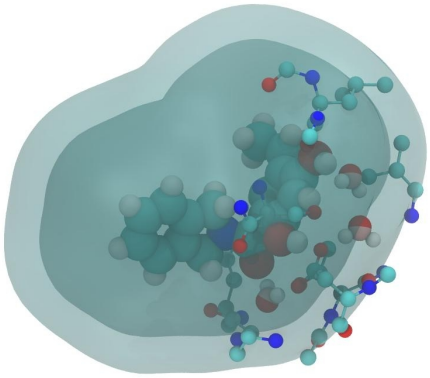


Figure 2.2: vdW method defines regions of grid points as function of distance from the vdW surface of a ligand. The grid region is contained within the opaque green surface, which typically can range to the second solvation shell but can be altered as desired.

2.3.2 Selecting by density clustering

One may also select grid points by density clustered regions. In this case the solvent densities of grid points can be utilised to generate a number of centroids around high density sites. This can be defined by having a minimum density cutoff (1.5 relative density to bulk) for a centroid centre and allocating neighbours within a cutoff distance from that cluster, typically 1.5 Å. From these centres clusters of interest which are near chemical moieties or connected to waters near them can be investigated with a spherical region. The overlap between the regions are taken to produce a union of all the grid regions of interest.

2.3.3 Selecting by solute vdW surface

Another simple method which is useful if full restraints on all heavy atoms of a ligand are being implemented is the vdW method, displayed in Figure 2.2. Here the coordinates of the minimised ligand conformation are taken, the appropriate vdW radius is then taken from the GAFF forcefield and used to define the initial vdW overlap of grid points which can be varied as a function of distance from the ligand. This is particularly effective at quickly converging the ligand desolvation cost and seems useful for capturing most solvent effects in ligand-protein simulations, though noise due to lack of sampling can always cause issues.

Chapter 3

Validation of GCT with bulk water and small molecule hydration studies

The chapter is mainly adapted from the paper, "Prediction of Small Molecule Hydration Thermodynamics with Grid Cell Theory" [80]. All the major calculations using GCT were computed by Georgios Gerogiokas (with the exception of some short calculations run on various small molecules with changed atom parameterisations done by Julien Michel for the discussion) with a TI implementation provided by Gaetano Calabro and run by Julien Michel. All other authors provided useful discussion and input but did not run calculations. The paper was used to validate application to biomolecular systems. This was done through comparison of predictions of hydration free energies to experimental values for different ions, noble gases, and aromatic molecules, as well as other polar and aliphatic small molecules which are similar to various environments found near different amino acid side chains and the protein backbones.

However, firstly the theory should also work in the simplest system, bulk water. Bulk water simulations are necessary to monitor the convergence of cell parameters at single grid points as well as entire regions of grid points. This was also important for identifying the balance between spatial resolution (grid density) and sampling time.

3.1 Molecular Models used for the study

Grid cell theory was used on a data set of small molecules including neon, xenon, chloride, sodium, methane, ethane, *n*-butane, isobutane, benzene, methanol, acetamide, and *n*-methylacetamide. For water only the TIP4P-Ew water model was used [81].

Noble gas parameters are derived from the work of Bondi [82] and Guillot and Guissani [83] and the ion parameters from work by Joung and Cheatham [18]. All other small molecules used the GAFF force field [44] and AM1-BCC charges [45] for parameterisation, as found in the AMBER11 software suite [84]. Unless otherwise stated, each small molecule was solvated in a box of 804 TIP4P-Ew water molecules with the program 'leap'. Afterwards the models were energy minimised and equilibrated under NPT conditions at 1 atm and 298 K. A velocity-verlet integrator and a time step of 2 fs were used. Temperature control used a Langevin thermostat with a coupling constant of 5 ps^{-1} [49], and the pressure was controlled using a Berendsen barostat with a coupling constant of 2 ps^{-1} [50]. All intramolecular degrees of freedoms in water molecules and bonds involving hydrogen atoms in solutes were constrained using the SHAKE method with a tolerance of 0.00001 Å [85, 86]. Electrostatic interactions were computed with particle mesh Ewald method [87, 88] with a cutoff of 10 Å. Lennard-Jones interactions were truncated at the same cutoff as the electrostatic interactions. The program 'sander' was used to run molecular simulations until the box density stabilised, which usually only requires about 100 ps. All solute heavy atoms were restrained to their initial conformation so that they were centred using harmonic positional restraints of force constant 10 kcal mol^{-1} $Å^{-2}$.

3.2 Molecular Dynamics Production runs

The software Sire/OpenMM was used to produce the molecular simulations for analysis. This program is created through a runtime linking of the general purpose molecular simulation package Sire (revision 1786) [89], with the GPU molecular dynamics library OpenMM (revision 3537). [90] Simulations were performed at 1 atm and 298 K with an atom-based generalized reaction field nonbonded cutoff of 10 Å for the electrostatic interactions [51], and an atom-based nonbonded cutoff of 10 Å for the Lennard-Jones interactions. Using harmonic positional restraints with a force constant of 10 kcal mol^{-1} $Å^{-2}$ on the solute heavy atoms, solutes were restrained to their input conformation. The temperature was controlled using an Andersen thermostat with a coupling constant of 10 ps^{-1} [48]. Pressure was controlled by attempting isotropic box edge scaling Monte Carlo moves every 25 time steps. The intramolecular degrees of freedom of water molecules and bonds involving hydrogen atoms were constrained using the OpenMM default error tolerance settings which is currently implemented in Sire. Each system was simulated three times for 50 ns using the same starting conformation but a different random velocity assignment, unless otherwise stated. This is to see the errors between replicates for the restrained conformation of interest. Every 1 ps a snapshot was stored in a DCD file format for subsequent analyses. The first 1 ns of sampling for each simulation were not considered in subsequent analyses.

3.3 *Nautilus* analyses

All simulations where analysed with the *Nautilus* software following the general protocol outlined in section A with the grid centred on the solute of interest. Cell files and then final grid files were then generated from the production run usually omitting the first nanosecond to be certain that the system was well equilibrated.

3.4 Thermodynamic Integration Calculations

A brief introduction into the thermodynamic integration calculations are outlined because the method was used as a comparison. The implementation for computation was made by Gaetano Calabro and calculations were done by Julien Michel. Absolute free energies of hydration were computed using a single topology coupling method [91], in the simulation package Sire/OpenMM (Sire revision 1994, OpenMM 5.1). The same potential energy function as that used in GCT was used, including solute restraints. Hydration free energies were computed in two stages. A total of 21 evenly spaced values of the coupling parameter λ (0.00, 0.05, ..., 0.95, 1.00) were used. A finite- difference thermodynamic integration (FDTI) approach was used to evaluate free energy gradients using a $\Delta\lambda$ set to 0.001 [92]. Numerical integration of the free energy gradients used a polynomial regression scheme [93]. The free energy change for turning off the atomic partial charges of the solute in vacuum was first computed ($\Delta G_{\text{vac\_coul\_off}}$). After that the free energy change for turning off the Lennard-Jones parameters of the discharged solutes in vacuum was then computed ($\Delta G_{\text{vac\_LJ\_off}}$). There then is a transfer of the solutes into a waterbox of 804 TIP4P-Ew water molecules, and the free energy change for turning on the Lennard-Jones interactions was computed ($\Delta G_{\text{solv\_LJ\_on}}$). Finally, the free energy change for restoring the atomic partial charges of the solutes in solution was computed ($\Delta G_{\text{solv\_coul\_on}}$). The hydration free energy is then

$$\Delta G_{\text{hyd}} = \Delta G_{\text{vac\_coul\_off}} + \Delta G_{\text{vac\_LJ\_off}} + \Delta G_{\text{solv\_LJ\_on}} + \Delta G_{\text{solv\_coul\_on}} \tag{3.1}$$

To avoid numerical instabilities, soft-core potential energy functions were used for transformations of the Lennard-Jones parameters [94, 95]. The implementation is identical to that used by Michel *et al.* [96]. The softening parameter was set to $\delta = 3.0$ except for waterbox simulations of ethane, benzene, isobutane, n-butane, and N-methylacetamide where it was set to 4.0 to avoid abrupt changes in free energy gradients which occurred when the Lennard-Jones interactions of the solutes is restored in the waterbox. The same input files were used for the waterbox FDTI simulations as for the GCT simulations. The same input files for the vacuum simulations were used as for solute conformation in the waterbox simulations. Each λ value was simulated for 1 ns (wa-

terbox) or 100 ps (vacuum). Solvent re-equilibration upon changes in λ was enabled by removing statistics from the first 100 ps of the waterbox simulations. The overall sampling time for the FDTI protocol was thus 42 ns waterbox and 4.2 ns vacuum. This is similar to the sampling time of a 50 ns GCT simulation in the current Sire/OpenMM implementation; the FDTI energy function is about 40% more computationally expensive to evaluate than the default MD energy function (the vacuum simulations have a negligible computing time). The various λ windows of the FDTI simulations can fortunately be run in parallel. Each hydration free energy calculation was run in triplicate using different random velocity assignments, and the mean and standard errors were computed. All calculations were performed on Tesla M2090 nodes using the OpenMM OpenCL platform in mixed precision mode.

3.5 Bulk Water

Bulk reference parameters for water molecules are needed to compare changes in thermodynamics of waters in different areas in the simulation space. In Table 3.1 these are all listed. The average forces, torques, orientational numbers, density and intermolecular energy per water molecule were averaged over triplicate 50 ns simulations of 804 water molecules. The results are similar to previously reported cell theory parameters for TIP4P-Ew, [61] though small differences are seen. This could be because reaction

Parameters and properties	TIP4P-Ew	Experiment[a]
$U^{inter}_{w(l)}$ (kcal mol^{-1})	-11.025(9)	-
$F^{1}_{w(l)}$ (10^{-10} N)	1.587(1)	-
$F^{2}_{w(l)}$ (10^{-10} N)	1.735(1)	-
$F^{3}_{w(l)}$ (10^{-10} N)	1.334(1)	-
$\tau^{1}_{w(l)}$ (10^{-20} N m^{-1})	1.061(1)	-
$\tau^{2}_{w(l)}$ (10^{-20} N m^{-1})	1.194(1)	-
$\tau^{3}_{w(l)}$ (10^{-20} N m^{-1})	1.453(1)	-
$\Omega^{ori}_{w(l)}$	3.305(4)	-
N^{bulk}_{w}	4.883(2)	-
ρ_b (kg m^{-3})	995.4(2)	997.1
ΔH (kcal mol^{-1})[b]	-9.98(1)	-9.92
ΔS (cal K^{-1} mol^{-1})[c]	-15.42(5)	-14.05

Table 3.1: [a]Reference [97] for ρ_b, ΔH, and ΔS. [b]Includes a dielectric depolarization correction term of 1.044 kcal mol^{-1} [81]. [c]Computed using eq. 7 from ref [61]. The dash (-) signifies not available. The numbers in the parentheses signify the standard error at the significant value provided.

field was used to treat long-range electrostatic interactions instead of particle mesh Ewald. Also, when different water models (SPC, SPC/E, TIP4P, TIP5P, TIP4P-EW and TIP3P) are considered there is little difference in bulk values which seems to signify little sensitivity to water models in cell theory [61]. In these cases forces and torques varied from each other by $\approx 1.5\%$ while energies differ by up to $\approx 4\%$. However, the results could be system dependent for more complex systems. For example, recent results in a study using GIST suggests that for Cucurbituril-guest binding, the TIP3P model gives more accurate results than TIP4P [16]. The excess enthalpies were also computed and then corrected for the intermolecular energy with the dielectric depolarisation correction term which occurs when water enters bulk [81]. Excess entropies were computed with eq. 7 of [61] and compared with experimental data in Table 3.1. The density and enthalpy of bulk water matched well, but the entropy is overestimated by 1% compared to experiment.

In this study the sampling required to obtain spatially resolved water properties is investigated. Grids made of voxels with sub-angstrom spacing enable fine spatial resolution but also require a larger number of snapshots to converge properties to an acceptable degree of precision, because each trajectory snapshot will contain on average fewer water molecules within each voxel k. Figure 3.1 shows the convergence of the enthalpies and entropies of water for evenly distributed regions $\mathbf{s}$ of space that altogether define a cube of volume of 4096 Å^3 centred at the centre of the box. Since the simulated system is isotropic, with sufficient sampling, the cell parameters of each region $\mathbf{s}$ should match the reference bulk parameters and the excess enthalpies and entropies should all converge to zero. In practice, insufficient sampling creates deviations. The distributions are approximately Gaussian and become more sharply peaked as the number of snapshots used for averaging increases. Results for other components are shown in Figure 3.2. Decreasing the grid resolution decreases the spread of all the distributions. For the components of entropy and water density in Figure 3.2 the distribution is Gaussian about zero as expected because zero indicates bulk-like properties. However there is noise in both water densities (Figure 3.2B) at very high resolution in the water orientational entropy due to sampling errors. In (Figure 3.2A) it can be seen that at especially low sampling (2000 snapshots) there is a discretisation error at high resolution (0.125 Å^3) because grid points require more sampling at those grid densities. However, the enthalpy distribution is not exactly centered on zero because the mean water intermolecular energy of the specific run used to produce Figure 3.1A differs slightly from the reference bulk value. This also shows the sensitivity of the calculation to the mean bulk water intermolecular energy. The entropy distribution on the other hand is almost exactly centered on zero. No systematic discretization errors are apparent. The results will converge to bulk properties as long as sufficient statistics have been collected for each region, regardless of the grid resolution. The entropy is consistently better converged than the enthalpy. For instance, with a 1 Å^3 spacing and

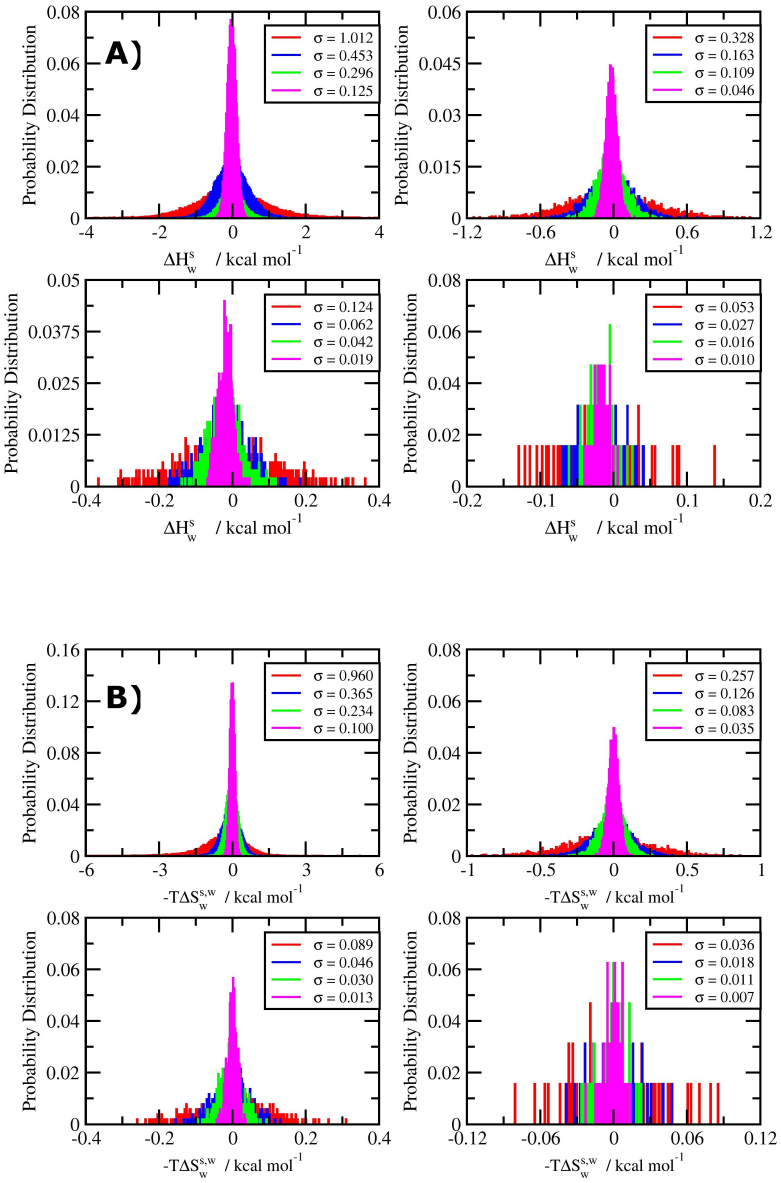


Figure 3.1: Distribution of (A) water enthalpies ΔH_{sw} and (B) water entropies $-T\Delta S_{s,ww}$ in bulk water. Each distribution was computed by dividing a cubic volume 4096 Å^3 centred at the centre of a box of 804 TIP4P-Ew molecules into evenly distributed regions of space **s** covering each 0.125 Å^3 (top left), 1 Å^3 (top right), 8 Å^3 (bottom left), and 64 Å^3 (bottom right). The red, blue, green, and magenta colors are distributions computed from a simulation of 2, 5, 10, and 50 ns duration, respectively. Data was only sampled after the first ns every 1 ps. The estimated standard deviation for each distribution using the full data set is shown in the legend.

50 ns averaging time, the standard deviation of $\Delta H_w^{\mathbf{s}}$ is larger than for $-T\Delta S_w$ ($\sigma =$ 0.045 kcal mol $^{-1}$ and $\sigma = 0.035$ kcal mol $^{-1}$ respectively).

Figure 3.3 illustrates the hydration enthalpy and entropy components of water within cubic regions **s** of increasing edge lengths.

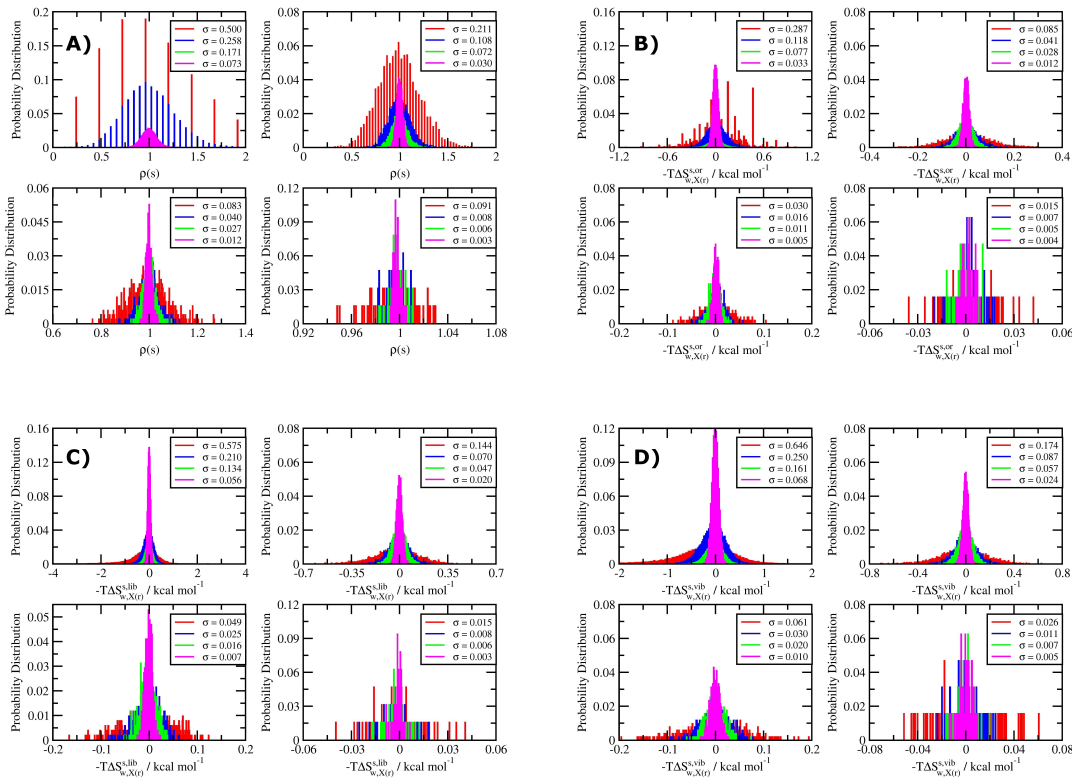


Figure 3.2: The distribution of A) water densities, B) water orientational entropies, C) water librational entropies, and D) water vibrational entropies in bulk water.Each distribution was computed by dividing a cubic volume 4096 Å^3 centred at the centre of a box of 804 TIP4P-Ew molecules into evenly distributed regions of space **s** covering each 0.125 Å^3 (top left), 1 Å^3 (top right), 8 Å^3 (bottom left) and 64 Å^3 (bottom right) respectively. The red, blue, green and magenta color indicate distributions computed from a simulation of 2 ns, 5 ns, 10 ns and 50 ns duration respectively. The first ns was discarded and snapshots were analysed every 1 ps. The legend indicates the estimated standard deviation for each distribution using the full dataset

Again, all results should converge to zero, but deviations will occur due to finite sampling errors. As expected, the uncertainty in the computed thermodynamic properties increases with the volume monitored by **s**. The enthalpy diverges more rapidly than the entropy components. The convergence behaviour of the enthalpy and entropy components can be rationalised by inspecting Table 3.1 and eqs. (2.13), (2.15) and (2.20) in chapter 2. As the volume of space covered by **s** increases, the number of water molecules N_w contributing to the enthalpy/entropy increases, and small random deviations of the averaged forces, torques and energies will contribute an increasingly significant enthalpy/entropy change. The enthalpy diverges more rapidly because more sampling is required to converge the mean per-water intermolecular energy and there is greater uncertainty in the value of the reference bulk parameter. For bulk water, a cube of edge length 12 Å and three simulations of 50 ns yield converged predictions to within $\pm$ 0.2 and $\pm$ 0.05 kcal mol^{-1} for the enthalpy and entropy respectively. The convergence behavior is likely to be system dependent; protein binding sites will typ-

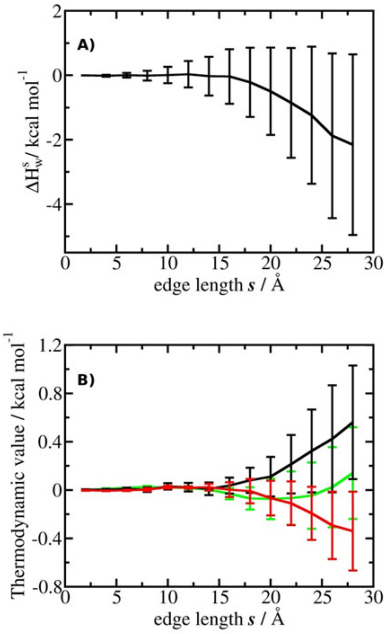


Figure 3.3: Convergence of the water (A) enthalpies $\Delta H_w^{\mathbf{s}}$ and (B) entropy components, $-T\Delta S_{w,X(r)}^{s,ori}$ (red), $-T\Delta S_{w,X(r)}^{s,vib}$(black), and $-T\Delta S_{w,X(r)}^{s,lib}$ (green) in bulk water as a function of the size of the monitored region **s**. Each data point is the mean of three 50 ns simulations, and the error bars show the standard error of the mean.

ically cover a smaller volume and contain fewer water molecules, which should lead to quicker convergence. On the other hand greater correlation times are expected for water molecules in the vicinity of biomolecular surfaces due to coupling. Nevertheless, the present results indicate that for quantitative studies, sampling errors introduce a limit to the volume of space **s** that can be reliably monitored with GCT.

3.6 Small Molecules

The enthalpies, entropies, and free energies of hydration of four monatomic solutes were computed. These simple solutes (neon, xenon, chloride and sodium), the properties of water molecules can be simply computed from grid volumes which are chosen as a function of the distance of the water oxygen atom to the centre of the solute. Also the the lack of solute conformational changes due to the absence of internal degrees of freedom makes it easier to compare with experimental data. To address the debate on the extent of solute perturbations on water structure, [98, 99] a detailed analysis of the computed enthalpies and entropies of hydration as a function of the volume monitored (defined by distance from the solute) was undertaken to establish up to what distances from a solute one should monitor solvent properties to observe convergence of free energies of hydration. Figure 3.4 presents the water enthalpy and entropy components

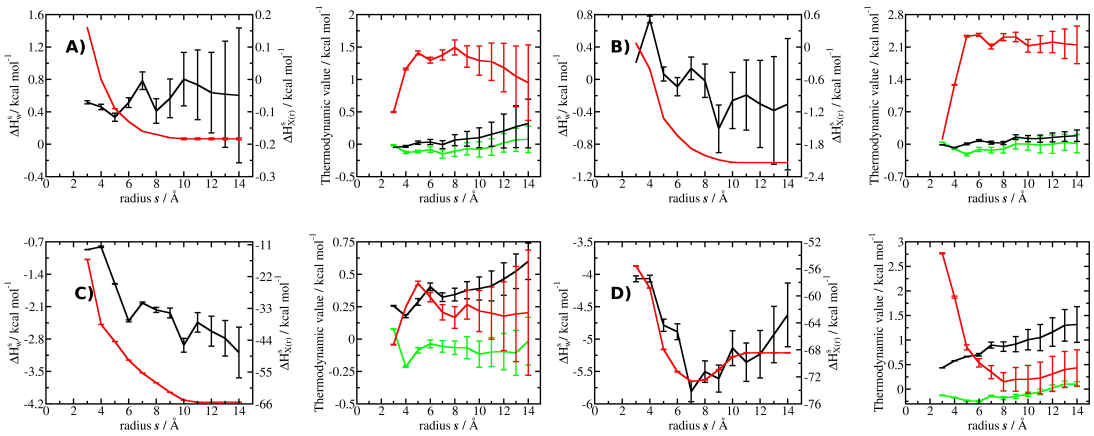


Figure 3.4: Convergence of hydration enthalpy and entropy components as a function of the size of **s** for (A) neon, (B) xenon, (C) Cl^{-1}, and (D) Na^{-1}. All solutes were solvated in a waterbox of 804 water molecules. The x-axis depicts the radius of a spherical region **s** centered on the solute. The y-axis indicates components of the enthalpy and entropy of hydration. Left panel: black ($\Delta H^{\mathbf{s}}_{w}$), red ($\Delta H^{\mathbf{s}}_{X(r)}$). Right panel: red ($-T\Delta S^{\mathbf{s},ori}_{w,X(r)}$), black ($-T\Delta S^{\mathbf{s},vib}_{w,X(r)}$), green ($-T\Delta S^{\mathbf{s},lib}_{w,X(r)}$). The error bars indicate the standard error of the mean obtained from three independent simulations.

as the radius of the monitored spherical region of space **s** centered on the solute atom is increased from 3 Å to 14 Å. In these systems, the box edge lengths fluctuate around 1-2 Å around 30 Å using periodic boundary conditions, so larger regions would include some water molecules twice.

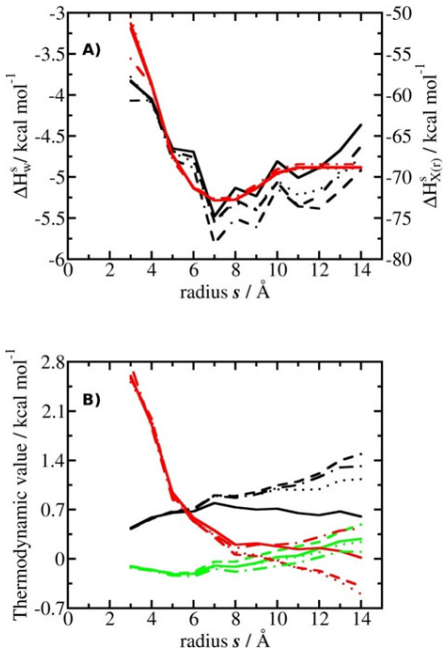


Figure 3.5: How box-size effects enthalpy A) and entropy B), coponents of the free energy of hydration of sodium. The solid, dashed and dotted, dashed, and dotted lines depict results for a box with average half-edge lengths of 15, 17.5, 20, and 25 Å, respectively. Error bars, which are comparable to those in Figure 3.4 have been omitted for clarity. Other symbols are as in Figure 3.4.

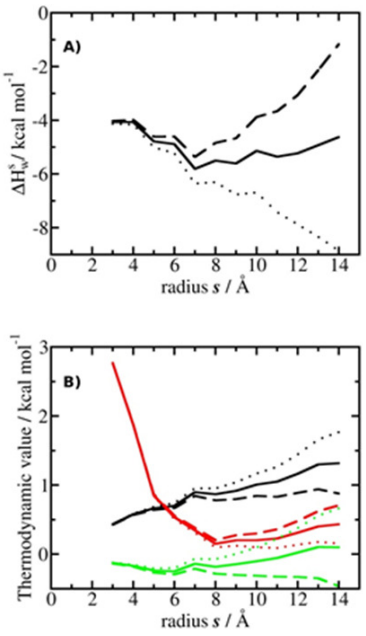


Figure 3.6: How uncertainties in the reference bulk parameters effect the enthalpy A), and entropy B), components of the free energy of hydration of sodium. The solid lines depict results obtained using the parameters listed in Table 1, and the dashed and dotted lines depict parameters modified by ± 1 standard error. Error bars, which are comparable to those in Figure 3.4 have been omitted for clarity. Other symbols are as in Figure 3.4.

The well converged solute-solvent enthalpy term $\Delta H^{\mathbf{s}}_{X(\mathbf{r})}$ fully accounts for radii greater than the 10 Å nonbonded cutoff used in the simulations. However, the solvent-solvent enthalpy term $\Delta H^{\mathbf{s}}_{w}$ is noisier, and the standard error of the mean is ± 0.3 kcal mol^{-1} for a radius of 10 Å and increases rapidly beyond this value. For the neutral solutes, this term appears to be reasonably flat beyond 10 Å, but it drifts upward for chloride or downward for sodium. These drifts may relate with the charged unit cell and finite-size effects due to the limited box size. The entropy components show similar trends but are typically more reproducible as seen in lower error between replicates. The orientational entropy is noisier than the other terms. In addition, the vibrational and librational entropy components show also a systematic, but less pronounced, drift beyond 10 Å radii similar to the solvent-solvent enthalpy term $\Delta H^{\mathbf{s}}_{w}$. It is interesting to compare with IFST studies of small molecule hydration. Huggins and Payne reported that with a spherical region of 12 Å radius centred on the solutes of interest, about 5,000,000 snapshots were required to converge entropies to within a decimal point, whereas only about 20,000 snapshots were needed to converge enthalpies to the same level of precision [100]. Alternative IFST implementations using a nearest-neighbor method instead of histograms may estimate entropy more efficiently [100]. A rigorous comparison would require analysis of identical trajectories of identical systems. Nevertheless, it appears that the GCT water entropy estimates converge faster than the IFST entropy estimates which requires smaller cell sizes of 0.5 Å. The origin of the systematic drifts in the

enthalpy or entropy components was explored using simulations of sodium in larger box sizes. The results shown in figure 3.5 show that the computed properties for radii up to 6-8 Å are well reproducible. Beyond that distance the components diverge with no trends with respect to box size, but systematic drifts remain. This suggests that the drifts cannot be explained by finite-size effects. Instead, their origin can be linked to uncertainties in the reference bulk parameter values. Figure 3.6 depicts how sensitive the GCT results are to variations of one standard error of the computed reference parameters. As the number of water molecules increases with the cube of the radial distance to the solute, the small systematic error in the reference properties of a bulk water molecule will scale causing large variations in the computed enthalpy/entropy components. The water-water enthalpy is particularly sensitive, as seen in the results depicted in figure 3.6.

At an 8 Å radius, there is an uncertainty of approximately 1 kcal mol^{-1} in ΔH_w, but the variability of the entropy components is only about 0.2 kcal mol^{-1} with a 10 Å radius. Therefore, if a volume of space **s** is monitored, greater systematic errors should be expected for computed absolute enthalpies than entropies of hydration but less so for relative enthalpies and entropies of hydration evaluated over the same volumes where this effect may largely cancel out. Also, the sensitivity of the computed properties to the grid density was assessed by performing analyses of the sodium simulations using a grid spacing of 0.5 or 2 Å. Little dependence on the grid spacing was observed (figure 3.7). Thus, for *Nautilus* analyses selection of an adequate voxel size should be primarily dictated by the desired trade-off between spatial resolution and trajectory size. So for protein-ligand systems, it appears reasonable to expect that a grid spacing of 0.5 or 1 Å will be sufficient; with a spacing of 0.5 Å, each voxel covers a volume of 0.125 Å^3, which amounts to 1/80th of the volume of a water molecule. The results here indicate that simulations on the order of 50 ns (50,000 snapshots) enable well-reproducible predictions of relative enthalpies and entropies of hydration by considering spherical regions centred on the solute of radius about 8-10 Å. However, if a greater volume must be considered, longer simulations should be performed. It should be emphasised that sampling errors scale with the number of water molecules, not the volume monitored. A sphere of 10 Å radius centered on sodium includes about 140 water molecules, which is much greater than the number of water molecules within a typical protein binding site (ca. 10-40).

Thus based on these considerations converged analyses of hydration properties of typical protein binding site appear feasible. The GCT enthalpy and entropy components in figure 3.4 gives insight into the breakdown of hydration of the monatomic solutes. The enthalpy of hydration of xenon (Figure 3.4B) is more negative than neon (Figure 3.4A) because of stronger Lennard-Jones interactions with water, which is slightly offset by a loss in water-water interactions. Water near the two solutes shows negligible changes in vibrational and librational entropies, but a significant difference in orientational en-

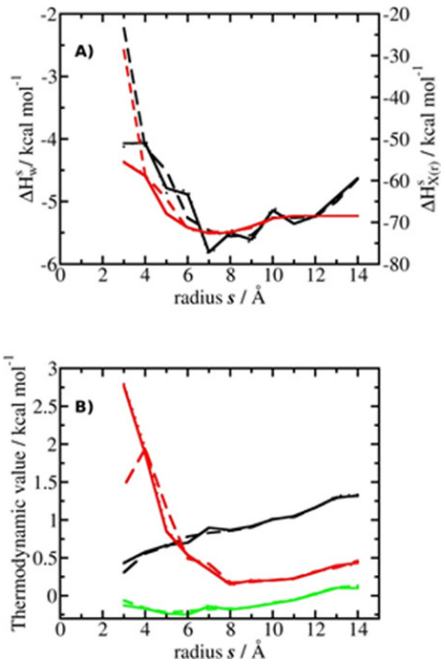


Figure 3.7: How grid spacing effects the computed enthalpy A), and entropy B), components of the free energy of hydration of sodium. The solid, dashed and dotted lines depict results for the box of 804 water molecules with a grid spacing of 1.0 Å, 0.5 Å and 2 Å respectively. Error bars, which are comparable to those seen in Figure 3.5 have been omitted for clarity. Other symbols are as in Figure 3.4 in the main text.

tropy due to the larger more hydrophobic xenon atom (but stronger Lennard-Jones interactions) which reduces more hydrogen-bonding arrangements of first and second shell water molecules. In xenon and neon the data converge at near the second solvation shell. Cl^{-1} (Figure 3.4C) has less negative solute-water enthalpy than sodium (figure 3.4D) but also a less negative water-water enthalpy component, so overall the enthalpy of hydration of sodium is more negative than chloride. For both ions the water orientational entropy component differs the most for first-shell water molecules because sodium cannot accept hydrogen bonds, but the changes in orientational entropy over larger volumes converge to similar values for the two ions. The water vibrational entropy component is lower for chloride than sodium, but the librational entropy component is small for both. The difference in vibrational entropy is probably due to stronger interactions of water with sodium for this force field, as shown by the enthalpy components in Figure 3.4C and D.

Next, *Nautilus* analyses were performed on eight neutral small molecules of varying polarity. Only water molecules within a 10 Å radius of the center of the solutes were considered to compute water enthalpy and entropy components because this cutoff gave acceptable statistical error and reproducible results. The full solute-water enthalpy term was used because it was converged and reproducible. Also, the solute translational and rotational entropies were evaluated using eqs. (1.31) and (1.32). FDTI (finite difference thermodynamic integration) was performed using Gaetano Calabro's

implementation in Sire/OpenMM. It was used to compute the hydration free energy of this set of small molecules using the same software and energy function as those in the *Nautilus* analyses. Figure 3.8A shows how free energies of hydration computed with FDTI and experimental data for the set of neutral molecules correlate. The results had a correlation coefficient of 0.99 and mean unsigned error of 0.98 kcal mol^{-1}, similar to those reported in the literature using similar force fields and methodologies [17, 101]. Chloride and sodium ions are not included in Figure 3.8A. Their computed free energies of hydration are -65.26 ± 0.04 and -69.34 ± 0.31 kcal mol^{-1}, respectively, which differ considerably from the experimental data of Schmid et al. [102], of -89.10 kcal mol^{-1} and -88.59 kcal mol^{-1}, respectively. This is because the results for the sodium and chloride ions were not corrected for systematic errors due to the use of a periodic boundary conditions and the reaction field treatment of long-ranged electrostatic interactions [103]. However, when appropriate corrections are made as derived from the work by Kastenholz and workers the estimate improves as follows. First a type A correction is made for the error in the solvent polarisation which occurs with the electrostatic cutoff. Kastenholz estimates a correction value of -38.54 kcal mol^{-1} for a sodium ion using a 10 Å cutoff, with an atomic cutoff for the generalised reaction field method in a system of 1024 SPC waters. There then is a type B correction which relates to the finite size of the box and its periodicity, where a correction term is -0.05 kcal mol^{-1} for the sodium ion. Finally, there is a type C correction term of -19.06 kcal mol^{-1} for sodium which corrects for the improper summation of the potential at the ionic site. However, as well as these terms you need a surface crossing term 16.8 kcal mol^{-1}. With all these terms there is an a large estimate for the sodium of -110.19 kcal mol^{-1}.

Figure 3.8B shows how the free energies of hydration computed with the two methodologies correlate. The GCT results strongly correlate with the FDTI results with a correlation coefficient of 0.97 and a mean unsigned error of 0.92 kcal mol^{-1}. The statistical error is higher in GCT calculations than the FDTI predictions, which is seen in the larger error bars in Figure 3.8B. This difference is due to the FDTI methodology which directly yields free energies of hydration through the evaluation of solute-solvent potential energy ensemble averages, on the other hand GCT free energies are obtained by summing enthalpies and entropies of hydration on each voxel that depend on slowly converging water-water energy terms. Note that one could evaluate entropies and enthalpies of hydration by thermodynamic integration, but the ensemble averages would also have noisy water-water terms of opposite magnitude that exactly compensate when the free energy of hydration is evaluated [104]. The correlation of the GCT computed free energies of hydration with experimental data for the neutral molecules is demonstrated in Figure 3.9A. The correlation coefficient is 0.98, and the mean unsigned error is 0.82 kcal mol^{-1}, showing higher accuracy than the FDTI predictions. As noted before, the GCT statistical error are larger, and FDTI is better suited for the evaluation of

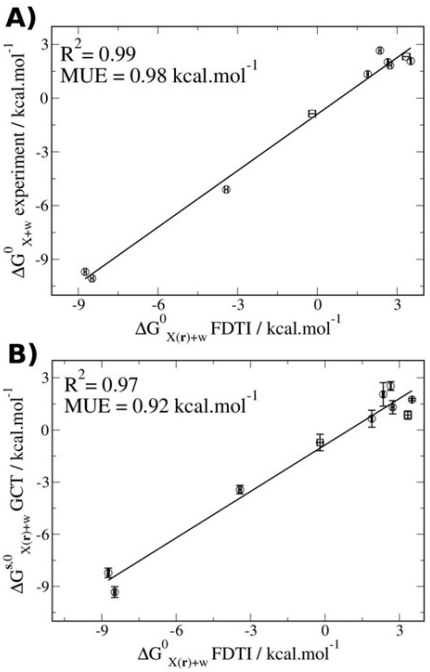


Figure 3.8: Accuracy of the thermodynamic integration predictions and correlation with grid cell theory results. (A) Correlation between finite difference thermodynamic integration free energies and experimental data. (B) Correlation between grid cell theory and finite difference thermodynamic integration free energies. The error bars show the standard error of the mean obtained from three independent simulations.

free energies of hydration for this data set. However, the *Nautilus* analyses provides enthalpies and entropies of hydration from a single simulation. These are plotted against experimental data in Figure 3.9B and C. There is some compensation in the systematic errors which is apparent in the higher mean unsigned error for both quantities which is about 1.3 kcal mol^{-1}, higher than for the free energies of hydration. The enthalpies of hydration and experiment correlate highly (R^2=0.98) but the correlation is much lower for the entropies of hydration (R^2=0.66). The lower correlation with the entropies of hydration could be due to the smaller energetic range of $-T\Delta S$ (about 7 kcal mol^{-1} versus 18 kcal mol^{-1} for the enthalpies) and some of the outliers.

The accuracy of the results is compared to a recent GIST study of Huggins and Payne [100] on the hydration thermodynamics of six small molecules. Although the smaller data set differs the dataset described in this work, the GCT enthalpies of hydration were similar to those reported by Huggins and Payne [100], but the GIST entropies of hydration were better correlated with experiment (R^2=0.77). Table 3.2 gives a comparison of calculated and measured enthalpies, entropies, and free energies of hydration for the data set of neutral and charged solutes. One sees that the free energies of hydration of neon and xenon are slightly underestimated, with greater discrepancies in the enthalpies and entropies of hydration. This is seen in previous results and suggests that the temperature dependence of the hydration of nonpolar solutes is

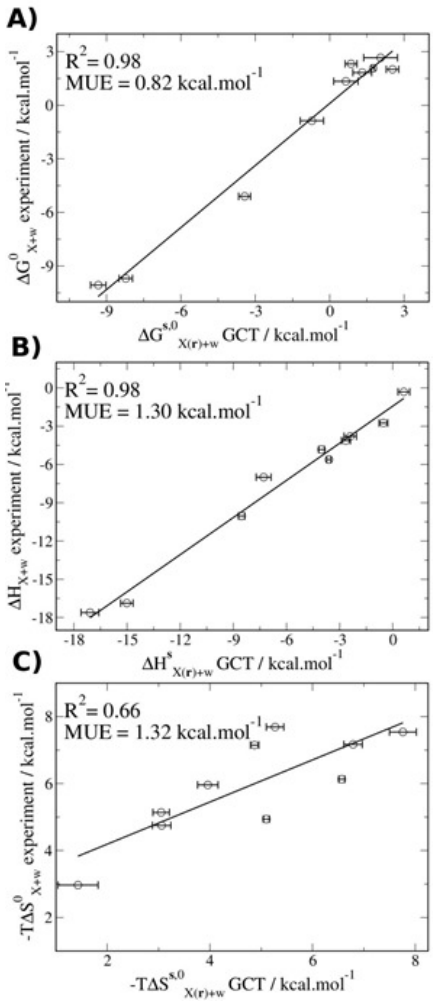


Figure 3.9: How grid cell theory computed hydration thermodynamics correlates with experimental data. (A) Correlation between computed free energies of hydration and experimental data. (B) Correlation between computed enthalpies of hydration and experimental data. (C) Correlation between computed entropies of hydration and experimental data. The error bars indicate the standard error of the mean obtained from three independent simulations.

not well captured with the present force field [63]. The effect is systematic, however the relative free energies, entropies, and enthalpies of hydration are in good agreement with experiment since there can be a cancellation of errors upon subtraction.

The computed properties for chloride and sodium are more challenging because finite size and cutoff errors are large for charged species. The problem with charged solutes has been explored in great detail by Hunenberger and co-workers [103, 105, 106] and correction terms for hydration free energies were computed using alchemical methods. More work would be required to derive correction terms for enthalpies and entropies of hydration computed using GCT. However, for the present study corrections cannot be derived for the ion since ionic Lennard-Jones and vdW parameters derived from Joung and Cheatham [18] are optimised for the ionic hydration free energy already. The present protocol results in sodium being better hydrated than chloride by about

4 kcal mol^{-1} which is not in agreement with the experimental data of Schmid et al. [102] listed in the table 3.2, but it should be noted that there is uncertainty in the experimental data owing to the difficulty of measuring the hydration thermodynamics of a proton [106]. However, there is a systematic small underestimation of the free energies of hydration for the other nonpolar solutes, ethane, isobutane, and n-butane. The magnitudes of both the enthalpies/entropies are also underestimated, which could again be due to the force field having issues capturing the temperature dependence of hydrophobic hydration. Additionally, for the largest outlier in the neutral molecules data set, isobutane, it is probable that the solute rotational entropy in solution is overestimated because of a weak interaction with water, which would also be reflected in the low computed torques [eq. (1.32)] yielding a null change in librational entropy. This reflects how the harmonic approximation of cell theory poorly captures entropy when interactions are weak.

For benzene, the enthalpies and entropies of hydration are well reproduced by GCT. The entropy of hydration of simulated methanol also correlates well with experiment, but the enthalpy is underestimated, which suggests that methanol is not sufficiently hydrated. This could be linked with an issue with the AM1-BCC / GAFF force field parameters used in the present study. The free energy of hydration of methanol computed with this force field but using different alchemical protocols also appears to be too positive by approximately 1-1.5 kcal mol^{-1} [17, 101]. This seems to be reflected in the parameterisation of the -OH group. Once the calculations were repeated after increasing the charge on the hydroxyl hydrogen atom by 0.05 e and decreasing the charge on the hydroxyl oxygen atom by 0.05 e; the enthalpy, entropy and free energy of hydrations were $\Delta H^{\mathbf{s},0}_{X(\mathbf{r})+w}$=-10.1 kcal mol^{-1}, -T$\Delta S^{\mathbf{s},0}_{X(\mathbf{r})+w}$= 5.3 kcal mol^{-1}, and $\Delta G^{\mathbf{s},0}_{X(\mathbf{r})+w}$=-4.8 kcal mol^{-1} respectively, in excellent agreement with experiment. The entropies of hydration of the more polar solutes such as acetamide and N-methylacetamide, are well reproduced by GCT. The free energies of hydration are 0.7-1.5 kcal mol^{-1} higher than experiment because the enthalpies of hydration are again not sufficiently negative. Similar to methanol, further polarization of the C-O or N-H bonds could address the error in the hydration enthalpies.

Overall, the results of both the predicted GCT enthalpies and entropies of hydration correlate well with experimental data. There are some issues caused by the forcefield used or the methodology itself which suggests there is room for further improvement. The biggest advantage of GCT is the ability to visualize the solvent regions of favourable and unfavourable contributions to the thermodynamics of hydration. The breakdown of free energy into enthalpic and entropic components elucidates the nature of the average local intermolecular interactions of water molecules. Figure 3.10 shows water's contributions to the free energy, enthalpy, and entropy of hydration of three small molecules, N-methylacetamide (NMA), methanol, and benzene. 5,000,000 snapshots were collected from 10 independent 50 ns simulations to finely resolve small energetic

	Computed			Experimental[b]		
	$\Delta H^{s}_{X(r)+w}$	$-T\Delta S^{s,0}_{X(r)+w}$	$\Delta G^{s,0}_{X(r)+w}$	ΔH_{X+w}	$-T\Delta S^{0}_{X+w}$	ΔG^{0}_{X+w}
Neon	0.62(33)	1.43(39)	2.05(68)	-0.31	2.97	2.66
Xenon	-2.40(35)	3.05(16)	0.65(49)	-3.80	5.14	1.34
Chloride	-68.36(15)	3.10(17)	-65.26(04)	-93.69	4.59	-89.10
Sodium	-73.61(29)	4.27(40)	-69.34(31)	-93.45	4.86	-88.59
Methane	-0.53(24)	3.06(18)	2.53(25)	-2.75	4.75	2.01
Ethane	-2.65(25)	3.96(20)	1.31(38)	-4.13	5.96	1.83
Isobutane	-4.02(17)	4.87(08)	0.86(24)	-4.83	7.15	2.32
N-butane	-3.61(12)	5.27(17)	1.75(06)	-5.62	7.69	2.07
Benzene	-7.29(42)	6.57(06)	-0.72(47)	-7.00	6.13	-0.87
Methanol	-8.53(18)	5.10(06)	-3.43(24)	-10.04	4.94	-5.10
Acetamide	-15.01(35)	6.79(18)	-8.22(27)	-16.87	7.17	-9.70
N-methylacetamide	-17.09(05)	7.76(26)	-9.33(31)	-17.61	7.54	-10.07

Table 3.2: a) All data are in kcal mol^{-1} b) Experimental data from Ref [107] for neon, xenon, Ref [102] for chloride, sodium. Ref [108] (ΔH_{X+w}) and Ref [109] (ΔG°_{X+w}) for acetamide and N-methylacetamide, Ref [110] for other solutes. Ref [110] gives enthalpies for a constant pressure solvation process px and these were converted using $\Delta H_{X+w} = \Delta H^{px}_{X+w} + k_B T$. When missing, entropies were derived from the difference of the Gibbs free energies with the enthalpies.

differences between neighboring voxels, and to produce grid cell files with cubic voxels of edge length 0.5 Å. The resulting contours are generally smooth, except when drawn at values close to zero (or one for the relative density) due to sampling errors. Harmonic restraints were applied to all solute atoms to avoid rotations of methyl hydrogen atoms or the polar hydrogen atom of methanol, which would blur the properties of nearby water molecules. For these solutes, the computed properties are similar and within statistical error of the results shown in Table 3.2. For NMA (Figure 3.10A), the highest contributing region to the free energy of hydration is found in a hemisphere about the amide oxygen atom. Water molecules there donate a hydrogen bond to the amide oxygen atom. Interestingly, configurations where a hydrogen bonding water introduces a linear angle between a water molecule oxygen atom, amide oxygen atom, and amide carbon atom are slightly less favored, possibly due to water network fluctuations.

A smaller contribution is obtained from a separate region where water molecules tend to accept a hydrogen bond from the amide polar hydrogen. Finally there is a third weakly stabilizing region around the methyl group bonded to the amide nitrogen atom. Water molecules here can favourably orient hydrogen atoms toward the amide oxygen

and nitrogen atoms, while minimizing electrostatic repulsion with the amide polar hydrogen. Regions above and below the amide bond plane contribute unfavorably to the free energy of hydration because of unfavorable enthalpic and entropic contributions, which were also identified with a GIST analysis by Huggins and Payne [100]. One also sees that water structuring around high density regions reduce water density further away from the solute, especially near the region where water molecules can accept a hydrogen bond from the solute. As expected water in high-density regions contributes the most to solvent entropy loss. Hydration entropy is less positive near the methyl groups of solutes. However, when there is a more positive entropy in the first solvation shell there is usually a small entropically favourable (more negative), regions observed in the second hydration shell compared to the first shell water molecules that are accepting or donating hydrogen bonds to the solute, an example of entropy/enthalpy compensation.

For methanol (Figure 3.10B) two distinct regions that contribute favorably to the free energy of hydration are seen. They are related to water molecules which in one case donate, and another, accept hydrogen bonds. The region where hydrogen-bond donation occurs covers a larger volume than the water hydrogen-bond accepting region because methanol accepts more hydrogen bonds than it donates, possibly due to either larger solvent accessible surface area or the fact that water has more donating groups. However, enthalpic interactions are stronger in the smaller hydrogen bond-accepting region, which is partly compensated by greater vibrational and orientational entropy loss within the water hydrogen bond-accepting region.

In the NMA case, both regions near the hydrogen-bond donor and hydrogen-bond acceptor contribute favourably to the enthalpy of hydration and unfavourably to the entropy of hydration, but the isocontours drawn at the same isovalues are smaller in extent, compared to methanol. For methanol small favourable entropic contributions and unfavourable enthalpic contributions in the second shell are apparent again showing enthalpy/entropy compensation. However, regions about the methyl group and perpendicular to the C-O-H plane have both unfavourable enthalpy and entropy contributions to the hydration free energy. The region above the hydroxyl group has a small favourable entropy and an unfavourable enthalpy. Overall, the contributions explain 5 kcal mol^{-1} more negative hydration free energy of NMA over methanol. In both NMA and methanol, the high water density regions ($\rho(\mathbf{s}) > 2.7$) correlate well with favourably contributing regions toward the hydration free energy. A complete match is not obvious in Figure 3.10 because the contours have been generated at different isovalues. It is important to realise that a highly localized water density is not systematically associated with a favourable contribution to the free energy of hydration; for instance, regions near the methyl group of methanol show higher densities than bulk ($\rho(\mathbf{s}) > 1.5$) but do not contribute favorably to the free energy of hydration (Figure 3.11).

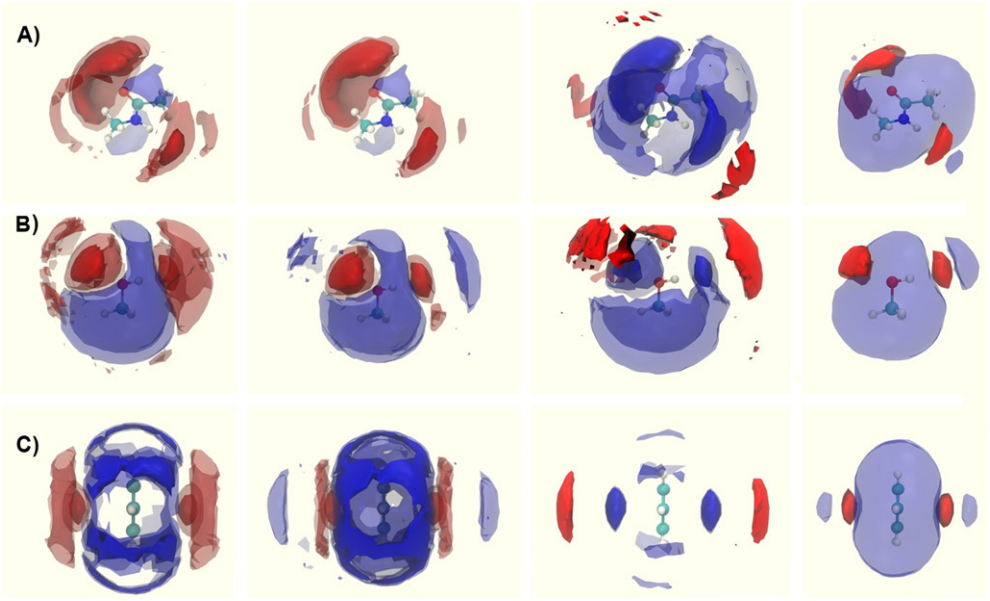


Figure 3.10: Spatial resolution of hydration thermodynamics around (A) N-methylacetamide, (B) methanol, and (C) benzene. For each solute, voxel contributions to $\Delta G_w^{\mathbf{s}}$, $\Delta H_{X(\mathbf{r})+w}^{\mathbf{s}}$, -T$\Delta S_w^{\mathbf{s},w}$, and $\rho(\mathbf{s})$ are shown from left to right. The blue isocontours indicate regions where water is less stable or has a lower density than in bulk. The red isocontours indicate regions where water is more stable or has a higher density than in bulk. The isocontours units are in kcal mol^{-1}Å^{-3} except relative density which is unitless. All density isocontours were drawn with the same isovalues $\rho(\mathbf{s})$: 2.7 (dark red); 0.4 (light blue). Other isovalues are (A) N-methylacetamide: $\Delta G_w^{\mathbf{s}}$; -0.1 (dark red), -0.0016 (light red), 0.02 (light blue). $\Delta H_{X(\mathbf{r})+w}^{\mathbf{s}}$; -0.1 (dark red), -0.03 (light red), 0.01 (light blue). -T$\Delta S_w^{\mathbf{s},w}$; -0.003 (dark red), 0.01 (light blue), 0.02 (dark blue). (B) Methanol: $\Delta G_w^{\mathbf{s}}$; -0.1 (dark red), -0.007 (light red), 0.05 (light blue). $\Delta H_{X(\mathbf{r})+w}^{\mathbf{s}}$; -0.1 (dark red), -0.02 (light red), 0.0025 (light blue). -T$\Delta S_w^{\mathbf{s},w}$; -0.002 (dark red), 0.006 (light blue), 0.05 (dark blue). (C) Benzene: $\Delta G_w^{\mathbf{s}}$; -0.1 (dark red), -0.009 (light red), 0.015 (light blue), 0.02 (dark blue). $\Delta H_{X(\mathbf{r})+w}^{\mathbf{s}}$; -0.1 (dark red), -0.019 (light red), 0.0025 (light blue), 0.008 (dark blue). -T$\Delta S_w^{\mathbf{s},w}$; -0.0027 (dark red), 0.015 (light blue), 0.02 (dark blue)

For benzene (Figure 3.10C), favourable contributions to the free energy of hydration arise from two small regions above and below the π-cloud (which is not explicitly modelled in the force field). Water molecules here weakly donate hydrogen bonds to the solute. In addition, two secondary doughnut-shaped regions provide additional weaker stabilizing enthalpic contributions. Water molecules here tend to donate hydrogen bonds to the water molecule(s) which are interacting with the π-cloud. Waters interacting in the plane of the ring are unfavourable to hydration, with greater losses in hydration free energy between two hydrogen atoms. This is because of an unfavourable enthalpic contribution, and a greater orientational entropy loss, because waters in these regions have fewer hydrogen bond acceptors in their coordination shell. Second shell waters are located above and below the plane of the benzene ring. They contribute slightly enthalpically unfavourable but have favourable entropy of hydration. The entropy gain is due to a favourable orientational entropy contribution whose effect is largely cancelled out by the enthalpy of hydration, with no significant contribution to

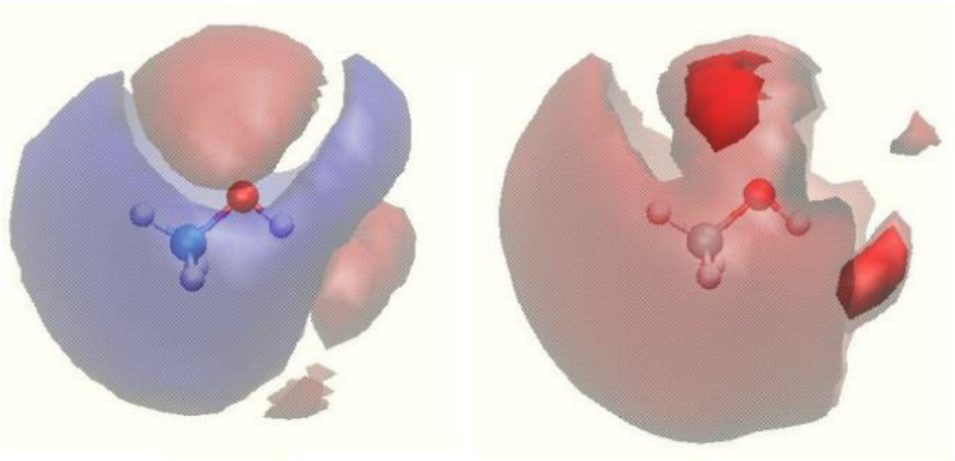


Figure 3.11: Components analysis of $\Delta G_w^{\mathbf{s}}$ (left) and $\rho(\mathbf{s})$ (right) for methanol. The same convention as in Fig 3.10 of the main text is used. The value of the isocontours are : -0.015 (light red), +0.005 (light blue) ; ($\rho(\mathbf{s})$) 3.0 (dark red), 1.5 (light red).

the free energy of hydration. All together, the results depicted in Figure 3.10 show that GCT provides a rich visualisation of hydration thermodynamics, which can yield insights into the localisation and nature of stabilizing/destabilizing solvent interactions.

3.7 Conclusions

Grid cell theory is a promising tool for molecular modelling studies of the hydration of organic molecules as well as biomolecules. However, if only free energies of hydration are of interest, alchemical free energy methodologies such as FEP and TI appear to be more reasonable options due to the slow convergence of water-water terms in GCT. On the otherhand a *Nautilus* analysis may be able to yield more insight into the thermodynamic decomposition of the free energy into the enthalpic and entropic driving forces. GCT also provides a visual component which gives insight into where favourable and unfavorable contributions are localised, enabling explanation of the overall free energy of hydration. This could then more adeptly aid molecular design of a ligand or solute in water. Also quantitatively, relative enthalpies and entropies of hydration are better converged with smaller uncertainties in the computed properties.

The most similar alternative to GCT is the GIST methodology proposed by Nguyen et al. [26]. Both approaches should provide similar enthalpies of hydration, essentially depending on the averaged interaction energies which are functions of the particular forcefield used. However, there are large differences in the entropy computation. GIST relies on an entropy expansion with multiple-particle correlation functions truncated at typically first order (sometimes second order) due to computational expense [26]. Er-

rors can arise from the neglect of entropic contributions from higher-order correlations. Cell theory uses a harmonic approximation which implicitly captures higher-order correlations in the cell parameters, and is used to describe vibrational and librational entropies, whereas the orientational entropy is captured with a generalized Pauling residual entropy model (there have been several improvements on the orientational entropy term from work by Henchman and Cockram [111]). Both methods seem to give results of similar accuracy for hydration free energies.

However, GCT has a number of practical advantages: the methodology appears less sensitive to the resolution of the grid and the solvent entropic components converge more rapidly than the enthalpic components. This could be a consequence of the functional form of the cell theory equations, i.e. entropies, are derived from logarithms of ratios, but enthalpies from differences of interaction energies. This should facilitate quicker routine applications since fewer snapshots must be post-processed from a MD simulation to attain a converged value. The computing time needed to perform a *Nautilus* analysis is a function of the size of the system simulated and the grid region. For example an analysis of 50,000 snapshots of a solvated small molecule in a cubic volume of 21,952 Å^3 currently requires ca. 40 CPU hours with the present implementation of the post-processing trajectory analysis software *Nautilus*. These can be easily chunked over hundreds of processors by processing trajectory snapshots concurrently. Further optimization is also possible by rewriting the software in a low-level programming language. Alternatively, grid properties could be computed "on the fly" through a parameter update step during a MD simulation.

Further work to assess how robust and accurate GCT predictions is needed with other solute/solvent force field combinations. Huggins has shown that IFST predictions between TIP4P-2005 and TIP5P-Ew, water models can cause a variations of up to 4 kcal mol^{-1} in the energy of the same hydration sites [112]. A rigorous comparison of GCT and IFST by analysis of the same MD trajectories with identical energy functions on larger datasets would be useful to assessing the merits of each approach and create better approaches to estimate solvent entropies. Irudayam and Henchman have also analysed the effect of different coordination environments of bulk water [98]. Their observations suggest that more detailed consideration of distributions of hydrogen-bonding configurations could improve estimations of solvent orientational entropies better than the Pauling-residual entropy model. Also since these results have been obtained for solutes restrained in a given conformation, further effects of flexibility could be investigated. GCT analyses can be performed on flexible solutes, but in its present implementation the grid would show blurry stabilised regions. However, if different conformational states are identified, one can perform separate analyses on conformation-specific grids [113]. Another idea is to use protocols that combine dynamical updating of the grid properties with a sliding time-window depending on conformation. Further research is needed to explore protocols to understand the coupling between solvent and

solute degrees of freedom.

The major strength of the GCT lies in its ability to spatially resolve enthalpy and entropy into physically intuitive components. It can also be used to guide the interpretation of free energy changes computed by alchemical methods, helping to rationalise results. On the basis of the results reported here further work was pursued on biomolecular systems to investigate how GCT could be used in a structure-based drug design context, and the results are detailed in chapters 4-6.

Chapter 4

Water displacement costs in scytalone dehydratase, p38 MAP kinase, and EGFR kinase systems

This chapter presents work that was published in [physical chemistry and chemical physics [114]]. One FEP/MD calculation was run by Stefano Bossisio. All other calculations were run by Georgios Gerogiokas with input into the interpretation of the work from the rest of the co-authors. The purpose of the work was to use GCT to give insights into the contributions of binding site water displacement on thermodynamics, which could be useful for ligand optimisation.

4.1 Importance of water displacement costs in the binding site

The effect of water on protein-ligand binding has been investigated because of its large contributions to the binding event in most cases. Research has suggested that binding-site water molecules are key components of the process [19–22]. The work done here explores perturbations of the binding-site water network structure and associated energetics that occur with small chemical modifications of small molecule ligands. This is an iterative task commonly tried during hit-to-lead and lead optimisation phases of a structure-based drug discovery campaign. Various computational methods are used to estimate water placement and energetics in binding-sites because individual water stabilities, and locations are not easily measured through experimental observables. Some commonly used software includes the following: the rolling probe-based GRID software [54]; molecular dynamics probes based methods such as MDMix, [115] MixMD, [116]

SILCS; [117] the Monte-Carlo lambda-dynamics based algorithm JAWS, [59,118] inhomogeneous fluid solvation theory (IFST) based techniques [26,29,30,53] including the popular method Watermap [119]; implicit and semi-explicit solvent methods such as SZMAP [67], three dimensional reference interaction site model (3D-RISM) [58], and variational implicit solvent model (VISM) [56].

The computational methods give insight into the role of water in protein-ligand interactions by elucidating water displacement costs in retrospective studies but have also been used in real structure-based medicinal chemistry efforts. For example the Watermap program has been used by Pfizer to gain insight into SAR behaviour which helped ligand optimisation of improved BACE-1 inhibitors [120].

Here work has gone into identifying a robust methodology with GCT for estimating reliable, accurate ranking of ligands. Also precision was desired, where the aim was to lower errors between replicates for the water displacement energetics in protein-binding sites. This is required to be able to discriminate how different ligands perturb binding-site waters. The approach relies on a discretisation of the cell theory method presented in chapter 3. The methodology has been validated by prediction of the hydration thermodynamics of small molecules [80] (see chapter 3), and has been applied to elucidate the binding thermodynamics of idealised host/guest systems [4]. The GCT method is applied in the present report for the first time to protein-ligand complexes in

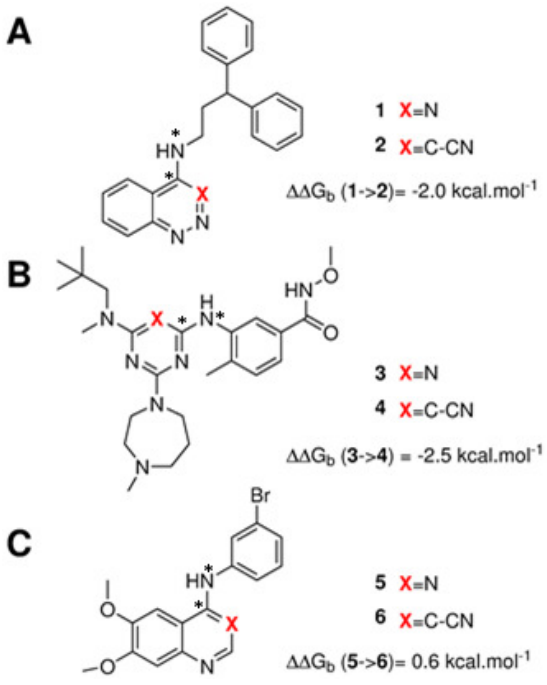


Figure 4.1: Structures of the three pairs of ligands studied. (A) Scytalone dehydratase, (B) p38 MAP kinase (C) EGFR kinase. Estimates of the experimental relative binding affinities are also shown. The star symbol denotes atoms used to define positional restraints (see section 4.2.2).

order to elucidate the impact of ligand modifications on the thermodynamic properties of binding site water molecules. Pairs of congeneric ligands of three different proteins were chosen for the work, Scytalone dehydratase, [121] p38 MAP Kinase, [122] and EGFR kinase [123]. In each case, a single binding site water molecule was displaced by introducing a cyano group and significant differences were observed in the changes in binding affinity (figure 4.1). Previous computational work has reproduced the observed trends in relative binding affinities with the use of Monte Carlo free energy perturbation methodologies (MC/FEP), but did not decompose the free energy into enthalpic and entropic components of the binding affinities or elucidate spatial details of the water network perturbations [124]. The objectives of the work were firstly, to assess whether GCT is a competitive alternative, secondly to determine solvent enthalpic and entropic contributions to binding affinities, and thirdly to determine the extent and nature of binding-site water perturbations upon ligand modification.

4.2 Theory and Method

4.2.1 Thermodynamic cycle

GCT analyses were performed using a thermodynamic cycle depicted in Figure 4.2. The free energy of water displacement from the thermodynamic cycle is given by eq.

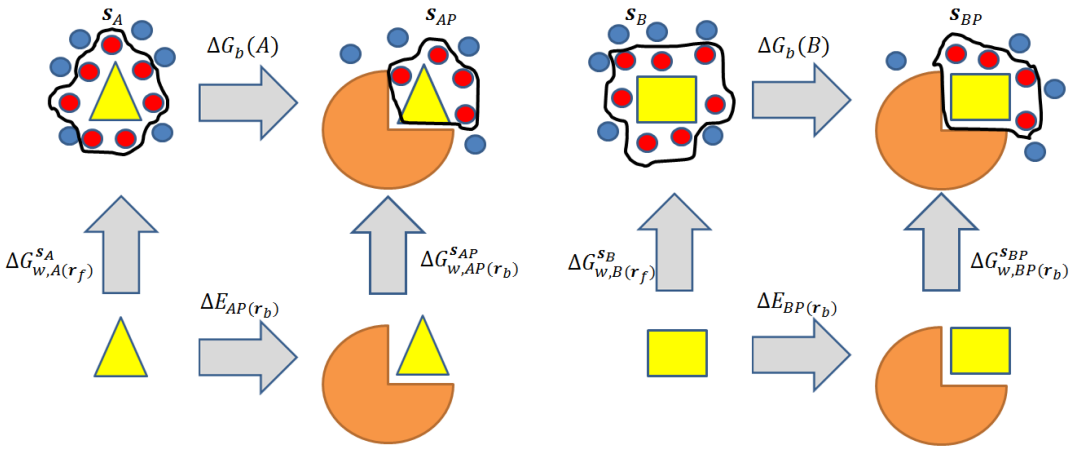


Figure 4.2: Thermodynamic cycles for evaluation of water displacement free energies and relative free energies of binding. Ligands are depicted by yellow shapes. Proteins are depicted by orange shapes. In all GCT analyses, water molecules (red circles) inside the monitored regions s_A, s_B, s_{AP}, s_{BP} contribute to the computed hydration free energies, whereas those that are out of the monitored regions are ignored (blue circles). Different restraint protocols r_c and r_l may be used to control allowed protein and ligand motions.

4.1:

$$\Delta\Delta G_{hyd}(AP \rightarrow BP, \boldsymbol{s}_{BP}, \boldsymbol{s}_{AP}, \boldsymbol{r}_c) = \Delta G^{\boldsymbol{s}_{BP}}_{w,BP(\boldsymbol{r}_c)} - \Delta G^{\boldsymbol{s}_{AP}}_{w,AP(\boldsymbol{r}_c)} \tag{4.1}$$

where $\Delta G^{\boldsymbol{s}_{AP}}_{w,AP(\mathbf{r}_c)}$ is the free energy of hydration of the region $\boldsymbol{s}_{AP}$ in the vicinity of ligand A and protein P using a restraint protocol $\boldsymbol{r}_c$. Ligand A is the ligand which does not displace a water but ligand B contains the cyano group which displaces the single water molecule. The water reorganisation energy is given by eq. 4.2.

$$\begin{aligned} &\Delta\Delta G_{water}(A \rightarrow B, \boldsymbol{s}_{BP}, \boldsymbol{s}_{AP}, \boldsymbol{s}_A, \boldsymbol{s}_B, \boldsymbol{r}_c, \boldsymbol{r}_l) \\ &= \Delta\Delta G_{hyd}(AP \rightarrow BP, \boldsymbol{s}_{BP}, \boldsymbol{s}_{AP}, \boldsymbol{r}_c) \\ &-\Delta\Delta G_{hyd}(A \rightarrow B, \boldsymbol{s}_B, \boldsymbol{s}_A, \boldsymbol{r}_l) \end{aligned} \tag{4.2}$$

There $\Delta\Delta G_{hyd}(A \rightarrow B, \boldsymbol{s}_B, \boldsymbol{s}_A, \boldsymbol{r}_l)$ is the difference between the hydration free energies of ligands A ($\Delta G^{\boldsymbol{s}_A}_{w,A(\boldsymbol{r}_l)}$) and B ($\Delta G^{\boldsymbol{s}_B}_{w,B(\boldsymbol{r}_l)}$) computed from grid regions $\boldsymbol{s}_A$ and $\boldsymbol{s}_B$ and restraint protocol $\boldsymbol{r}_l$. Then the relative binding affinities can be computed with eq. 4.3:

$$\begin{aligned} &\Delta\Delta G_b(A \rightarrow B, \boldsymbol{s}_{BP}, \boldsymbol{s}_{AP}, \boldsymbol{s}_A, \boldsymbol{s}_B, \boldsymbol{r}_c, \boldsymbol{r}_l) \\ &= \Delta\Delta G_{water}(A \rightarrow B, \boldsymbol{s}_{BP}, \boldsymbol{s}_{AP}, \boldsymbol{s}_A, \boldsymbol{s}_B, \boldsymbol{r}_c, \boldsymbol{r}_l) \\ &+\Delta\Delta E(AP \rightarrow BP, \boldsymbol{r}_c) \end{aligned} \tag{4.3}$$

where $\Delta\Delta E(AP \rightarrow BP, \boldsymbol{r}_c)$ is the interaction energy difference of ligand A ($\Delta E_{AP}(\boldsymbol{r}_c)$) and ligand B ($\Delta E_{BP}(\boldsymbol{r}_c)$) with protein P. Again as previously stated in chapter 2 contributions from relative changes in the ligand's strain energy (internal energies), translational/rotational entropies, and ligand-protein conformational entropies are neglected in the thermodynamic cycle used. The approximation should be reasonable between congeneric ligands which adopt similar binding modes.

4.2.2 Restraint protocols

GCT calculations were performed with several different protocols that vary in their use of restraints to control the conformations sampled by the ligands or protein during the simulations. GCT calculations can in principle be performed without any restraints on solutes; however this has a number of disadvantages. Firstly, extensive conformational sampling is required to obtain converged water properties for flexible solutes. Secondly, graphical analyses of voxel properties are more complex. Thirdly, the thermodynamic cycle depicted in figure 4.2 does not consider contributions from changes in conformations or flexibility from the protein and ligands. On the other hand restraints are artificial and may negatively affect the predictions of free energies of binding. In the present work different restraining protocols $\mathbf{r}$ were compared in an effort to identify a practical protocol for routine calculations. In the $\mathbf{r} = rot$ protocol positional restraints were applied on two atoms of a ligand. This was done to suppress rigid body motions

of the ligand. For all ligands the restrained atom is denoted by a star in figure 4.1. In the $\mathbf{r} =$ *bb* protocol, positional restraints were applied to protein backbone heavy atoms only. Finally, in the $\mathbf{r} =$ *full* protocol, positional restraints were applied to all heavy atoms of both ligand and protein. When in solution with the *full* protocol, ligands were restrained in their binding site conformation. Restraints were implemented with a force constant of 10 kcal mol^{-1} Å^{-2} for the *bb* and *full* protocols, and with a force constant of 5 kcal mol^{-1} Å^{-2} for the *rot* protocol. All restraints were applied on absolute Cartesian coordinates.

4.2.3 Preparation of Molecular Models

Models of scytalone dehydratase in complex with ligands **1** and **2** were generated using the PDB structure 3STD which was in complex with **2**. The crystal structure of EGFR kinase in complex with erlotinib (PDB 1M17) was used to define the binding mode of **5** and **6**. For p38a MAP kinase case the crystal structure of PDB 1DI9 which is in complex with a quinazoline inhibitor, was used to generate the protein model. AutoDock Vina [125] with the pymol plugin [126] was used to find suitable binding modes for **3** and **4** that matched structural data reported by Liu et al. [122] for **4**. The lowest energy pose produced by Vina for **4** was found to bind in a similar orientation. The TIP4P-Ew water model was used throughout. [81] All the small molecules were parameterized using the GAFF force field [44] and AM1-BCC charges [45], as implemented in the AMBER11 software suite [84]. For the protein, the ff12SB force field was used. Each protein complex and ligand was solvated with water extending 12 Å away from the edge of the solutes before performing energy minimisation. The preparation of molecular models was largely automated by the use of the software FESetup [127].

4.2.4 Molecular dynamics simulations

Molecular simulations were produced using the software Sire/OpenMM, which in the present study results from the linking of the general purpose molecular simulation package Sire (revision 1786), with the GPU molecular dynamics library OpenMM (revision 3537) [89]. Simulations were run at 1 atm and 298 K using an atom-based generalized reaction field nonbonded cutoff of 10 Å for the electrostatic interactions [51], and an atom-based nonbonded cutoff of 10 Å for the Lennard-Jones interactions. A velocity-Verlet integrator with a time step of 2 fs was used. Temperature control was achieved with an Andersen thermostat with a coupling constant of 10 ps^{-1}. [48] Pressure control used attempted isotropic box edge scaling Monte Carlo moves every 25 time steps. The OpenMM default error tolerance settings were to constrain the intramolecular degrees of freedom of water molecules. For each system triplicate simulations of 22 ns were run using the identical starting conformation for each run but initiated with a different

random velocity assignment. Snapshots were stored every 1 ps and were written into a DCD format. The first 1 ns of each trajectory was discarded to enable equilibration.

4.2.5 Grid cell theory analyses

All GCT analyses were performed with the trajectory post-processing software Nautilus (see section A). Bulk parameters for TIP4P-Ew were taken from a previous GCT study [80], which is also shown in Table 3.1 of chapter 3. The following protocols were used to define the regions of space subjected to Nautilus analyses. For each simulation, a 3D grid of evenly spaced points was centered on the Cartesian coordinates of the centre of mass of the ligand (x_{com},y_{com}, z_{com}). A rectangular region with minimum and maximum coordinates ($x_{com} \pm \Delta x$, $y_{com} \pm \Delta y$, $z_{com} \pm \Delta z$) was next defined and filled with grid points spaced every 0.5 Å along the x, y, and z components. The parameters Δx, Δy and Δz were chosen such that the grid would extend well beyond the ligand atoms or binding site region of interest (typical values are 11-14 Å). Cell parameters for every grid point within this rectangular region were then computed. For the simulations of the unbound ligands with the restraint protocol $\mathbf{r} = \textit{full}$, regions $\boldsymbol{s}_A/\boldsymbol{s}_B$ were defined as the union of the set of grid points that were within X_{vdw} Å of the van-der-Waals surface of the ligands A or B respectively. AMBER GAFF forcefield radii were used to define the van-der-Waals surface from the input ligand coordinates and several values of X_{vdW} were tested. For the simulations of the unbound ligands with the protocol $\mathbf{r} = \textit{rot}$, regions $\boldsymbol{s}_A/\boldsymbol{s}_B$ were defined as the length X_{cubic} of the edge of a cube centred on (x_{com},y_{com}, z_{com}).

For the simulations of the bound ligands with the restraint protocol $\mathbf{r} = \textit{full}$ or $\mathbf{r} = \textit{bb}$, regions $\boldsymbol{s}_{AP}$ / $\boldsymbol{s}_{BP}$ were defined by density-clustering of the trajectories of AP and BP. All grid points were first sorted by their local water density $\rho(k)$ in the simulation of AP. The medoid of a cluster was taken to be the grid point with highest density, and all grid points within 1.5 Å of this medoid were assigned to the cluster. All grid points belonging to the cluster were then removed from the grid and the process was iterated until no grid point with a density greater than 1.5 times bulk was found, yielding k_{AP} medoids. The process was repeated for the trajectory of BP, yielding k_{BP} medoids. Next, only medoids present in the binding site region of interest were retained; these were typically medoids present in binding site regions disconnected from bulk. Next, regions $\boldsymbol{s}_{AP}$ or $\boldsymbol{s}_{BP}$ were defined by selecting all grid points within X_{medoid} Å of each of the k_{AP} and k_{BP} medoids. In some instances, the medoids from the AP or BP simulations had very similar coordinates and a single medoid was retained. The procedure yielded a monitoring region $\boldsymbol{s}_C$ that is the union of $\boldsymbol{s}_{AP}$ and $\boldsymbol{s}_{BP}$. In some instances, additional analyses were performed by breaking-down $\boldsymbol{s}_{AP}$ or $\boldsymbol{s}_{BP}$ into M sub-regions $\{ \boldsymbol{s}_0, \ldots , \boldsymbol{s}_m \}$. This was done by defining a centre $\boldsymbol{r}_m = (x_m,y_m,z_m)$ for each of the M regions. The distance d_{im} of each grid point i in $\boldsymbol{s}_{AP}$ /$\boldsymbol{s}_{BP}$ to each $\boldsymbol{r}_m$

was computed and the grid point was assigned to the region M with the smallest value of d_{im}.

4.3 Results

4.3.1 Ligand hydration energetics

Figure 4.3A shows that the computed free energy of hydration of a ligand in the GCT formalism depends on the size of the monitored region, which are also shown in Table 4.1 on rows 3 and 4. With the $\boldsymbol{r} = \textit{full}$ restraint protocol, hydration free energies have converged for regions that extend approximately 6 Å away from the van der Waals surface of the ligands. For regions of this size, uncertainties in the absolute hydration free energies are on the order of 1 kcal mol^{-1}.

As discussed previously, GCT hydration free energies are less precise than those computed by FEP or TI approaches because of the contribution of water-water interaction energies to the enthalpy of hydration (Figure 4.3B). The entropies of hydration (figure 4.3C) are by contrast typically slightly better converged [63]. For the purpose of computing relative binding free energies between a pair of ligands the relative free energies of hydration are computed. Figure 4.3A shows that reasonable estimates of the relative free energy hydration of ligand differences can be estimated with smaller regions that extend to about 4 Å away from the van der Waals surface of the ligands. The use of very large GCT regions is actually detrimental to accuracy since the magnitude of uncertainties increases with the size of the region due to sensitivity to bulk parameters of the water. Overall for these ligands, a good trade-off is to select a value of the parameter X_{vdW} between 4-6 Å. Fig. 4.4 shows the computed hydration free energies (figure 4.4A), hydration enthalpies (figure 4.4B), and hydration entropies (figure 4.4C) with the $\boldsymbol{r}_l = \textit{rot}$ protocol. Convergence of the computed hydration energetics is observed for cubes of edge length X_{cubic} ca. 10 Å. The uncertainties in the computed quantities are larger than with $\boldsymbol{r}_l = \textit{full}$ protocol since the volume of the monitored region is actually larger. The relative hydration free energies are broadly comparable between the two restraint protocols for ligands **1**, **2** and for ligands **3**, **4**, but a noticeable discrepancy is apparent for ligands **5**, **6** ($\Delta\Delta G_{hyd}$(**5** $\rightarrow$ **6**, $\mathbf{s}_6$, $\mathbf{s}_5$, $\boldsymbol{r}_l = \textit{full}$, X_{vdW} = 6 Å) = -7.3 $\pm$ 0.5 kcal mol^{-1}, versus $\Delta\Delta G_{hyd}$(**5** $\rightarrow$ **6**, $\mathbf{s}_6$, $\mathbf{s}_5$, $\boldsymbol{r}_l = \textit{rot}$, X_{cubic} = 9 Å) = -0.3 $\pm$ 1.5 kcal mol^{-1}).

Visualisation of the trajectories indicates that this likely occurred because the pyrimidine N1 nitrogen of **5** is poorly hydrated owing to the close proximity of the bromophenyl group in the $\boldsymbol{r}_l = \textit{full}$ simulations. This occurred because the ligand was restrained to adopt the binding mode seen in the complex with EGFR kinase. Without

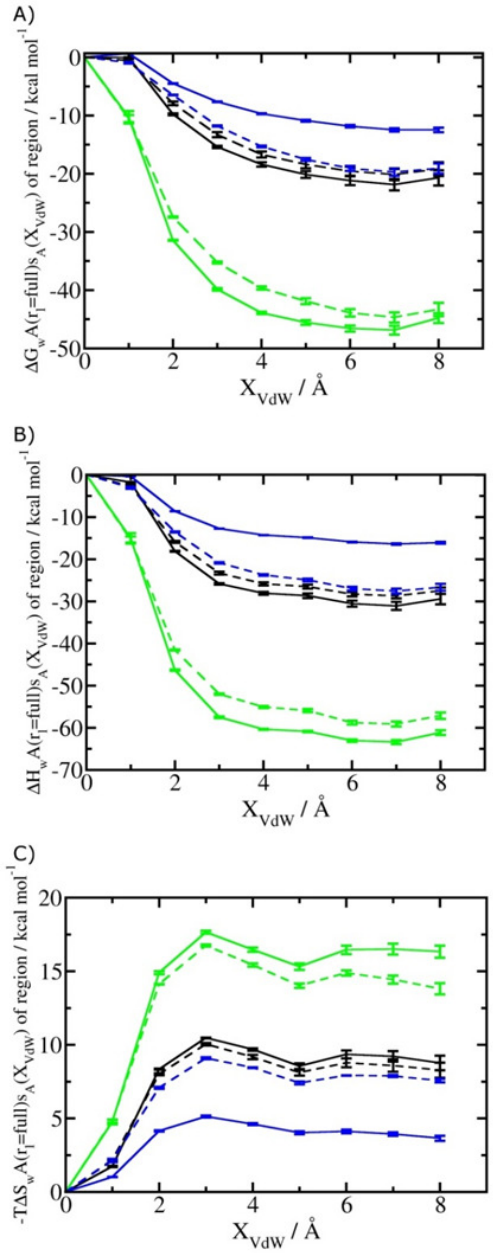


Figure 4.3: Ligand hydration energetics with the *full* restraints protocol (A) $\Delta G^{\boldsymbol{s}_A}_{w,A(\boldsymbol{r}_l)}$, (B) hydration enthalpies $\Delta H^{\boldsymbol{s}_A}_{w,A(\boldsymbol{r}_l)}$ and (C) hydration entropies $-T\Delta S^{\boldsymbol{s}_A}_{w,A(\boldsymbol{r}_l)}$ black lines are for the scytalone dehydratase ligands **1** (solid) and **2** (dashed). Green lines are for the p38 MAP kinase ligands **3** (solid) and **4** (dashed). Blue lines are for the EGFR kinase ligands **5** (solid) and **6** (dashed). Error bars represent the standard error of the mean computed from triplicate independent simulations.

such restraints in the $\boldsymbol{r}_l = rot$ simulations, **5** relaxed to a different conformation that increases hydration of the pyrimidine N1 nitrogen in **5**.

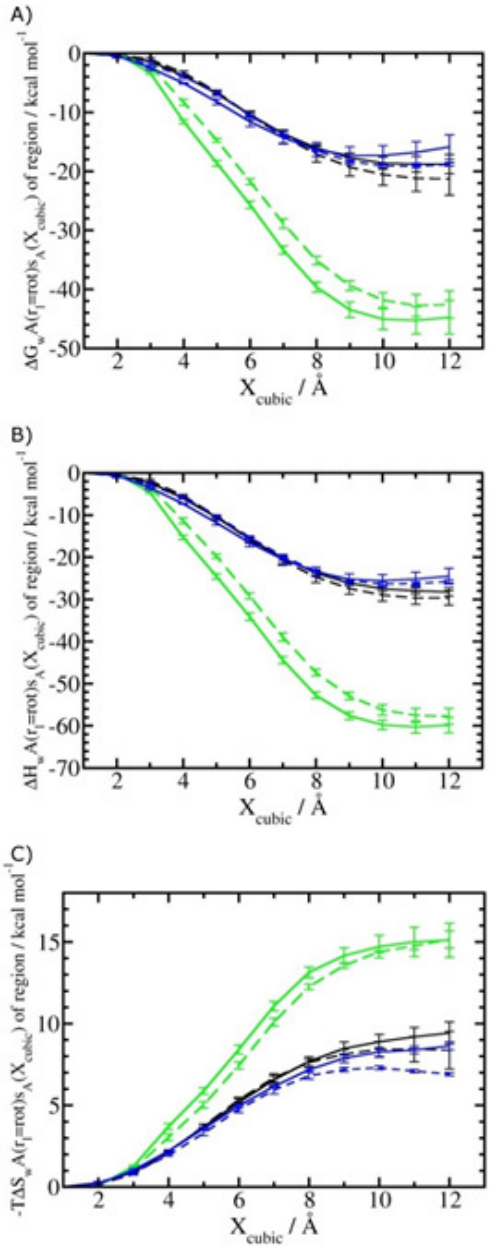


Figure 4.4: Ligand hydration energetics with the rigid body rotation restraints protocol (A) $\Delta G^{\boldsymbol{s}_A}_{w,A(\boldsymbol{r}_l)}$, (B) hydration enthalpies $\Delta H^{\boldsymbol{s}_A}_{w,A(\boldsymbol{r}_l)}$ and (C) hydration entropies $-T\Delta S^{\boldsymbol{s}_A}_{w,A(\boldsymbol{r}_l)}$ black lines are for the scytalone dehydratase ligands **1** (solid) and **2** (dashed). Green lines are for the p38 MAP kinase ligands **3** (solid) and **4** (dashed). Blue lines are for the EGFR kinase ligands **5** (solid) and **6** (dashed). Error bars represent the standard error of the mean computed from triplicate independent simulations.

4.3.2 Protein-ligand complex hydration energetics

Figure 4.5A shows the convergence of $\Delta\Delta G_{hyd}$(AP $\rightarrow$ BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$) as a function of time, for three different monitored regions $\boldsymbol{s}_c$ defined by varying the parameter X_{medoid}, and for two different restraining protocols $\boldsymbol{r}_c$. For low or intermediate values of X_{medoid}, similar results are obtained and trajectories of ca. 15 ns are needed to

observe convergence. The same hydration free energy is obtained because the larger region defined with X_{medoid} = 4 Å still includes only one water molecule. However for X_{medoid} = 8 Å, the hydration free energies differ markedly because the monitored region is now sufficiently large that it includes additional water molecules, some of them located out of the binding site of scytalone dehydratase. The hydration energetics are therefore different, and in the case of the $\boldsymbol{r}_c = bb$ protocol, no convergence is observed. The hydration energies between the $\boldsymbol{r}_c = bb$ and $\boldsymbol{r}_c = full$ protocols are not consistent because conformational changes in protein residues during the $\boldsymbol{r}_c = bb$ simulations affect the energetics of water molecules within the monitored region $\boldsymbol{s}_c$.

Figure 4.5B shows the computed hydration energetics for the three complexes using the *full* trajectories, but varying X_{medoid}. The plots show that the changes in hydration energetics between ligand pairs are relatively constant for small values of X_{medoid}, and the $\boldsymbol{r}_c = full$ protocol. Larger fluctuations are seen for EGFR kinase since the monitoring region $\boldsymbol{s}_c$ is larger and contains more water molecules. Larger values of X_{medoid}, or the additional protein flexibility in the $\boldsymbol{r}_c = bb$ protocol, causes increased statistical errors and fluctuations in the computed energetics. This indicates that much longer trajectories would be needed to obtain reproducible changes in hydration energetics of the complexes. Consequently similar variability is seen in the evaluation of water reorganisation energies with eq. 4.2 (figure 4.5C). Overall, with trajectories of the order of ca. 10 ns, it seems advisable to use the $\boldsymbol{r}_c = full$ protocol with X_{medoid} values between 4 to 6 Å if reproducible hydration energies are desired. Figure 4.6A shows the water content of the monitored region for the scytalone dehydratase/**1** complex. A single buried water molecule is present, hydrogen-bonded to two nearby tyrosine side-chains, and the nitrogen N1 of **1**. As expected, the water molecule is displaced in the scytalone dehydratase/**2** complex, and the cyano group is instead hydrogen-bonded to the two tyrosine phenolic hydroxyl groups (figure 4.6B). Note, here a hydrogen bond is simply defined if two electronegative atoms, with one atom being bonded to a hydrogen, is within 3.2 Å of each other. The monitored region in the p38 MAP kinase/**3** complex contains two water molecules that mediate hydrogen-bonding interactions between the ligand and the protein (figure 4.7A). Interestingly, the $\boldsymbol{r}_c = full$ and $\boldsymbol{r}_c = bb$ protocols lead to qualitatively different monitored regions.

This is because in simulations of the complex with **3** under $\boldsymbol{r}_c = bb$ conditions, one of the two water molecules may sometimes migrate to a third position, and then escape from the binding site. This occurred in ca. 5 ns in the first replicate, but did not occur in the second replicate, and occurred after 3 ns in the third replicate, but another water molecule returned after 20 ns to reproduce the original hydration state. Ideally a $\boldsymbol{r}_c = rot$ simulation would be run but the sampling would be larger. This suggests a slow equilibrium between at least two hydration states. By contrast, the picture that emerges from simulation of **4** with the two restraining protocols is relatively consistent (figure 4.7B). In the case of EGFR kinase in complex with **5** (figure 4.8A), the monitored

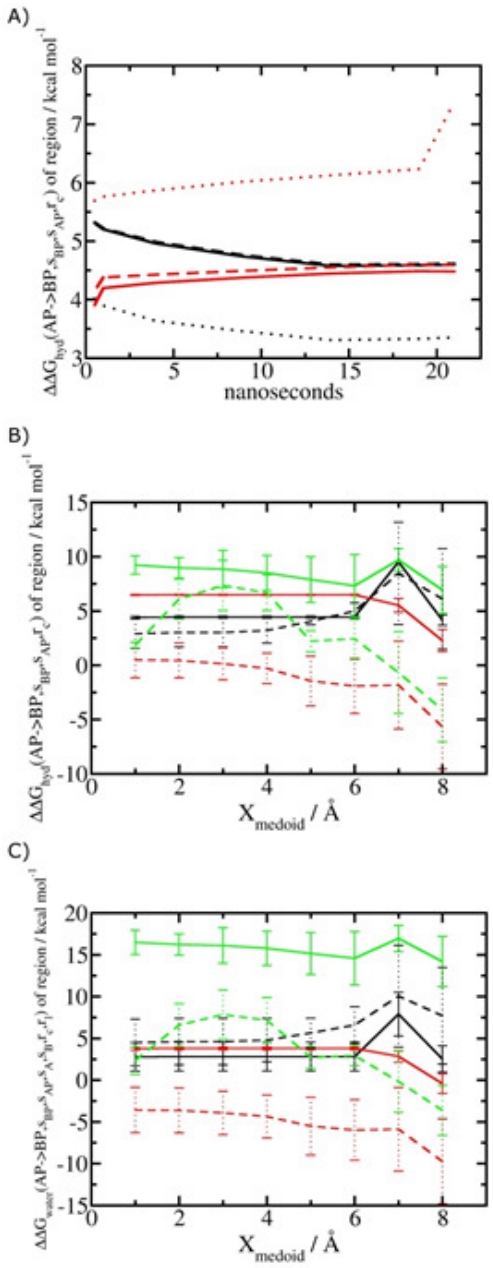


Figure 4.5: Convergence of hydration energetics and water reorganisation energetics for protein-ligand complexes. (A) Convergence of hydration energetics eq. 4.1 with respect to trajectory duration for scytalone dehydratase. Results in black are for $\boldsymbol{r}_c$ = *full*, and in red for $\boldsymbol{r}_c$ = *bb*. The solid line is for X_{medoid} = 1 Å, the dashed line for X_{medoid} = 4 Å, and the dotted line for X_{medoid} = 8 Å. (B) Hydration energetics as a function of Xmedoid using the *full* trajectories for scytalone dehydratase (black) p38 MAP kinase (red), and EGFR kinase (green). Solid lines are the results obtained with the $\boldsymbol{r}_c$ = *full* protocol and dotted lines are the results obtained with the $\boldsymbol{r}_c$ = *bb* protocol. (C) Same as (B) but for the water reorganisation energy (eqn 4.2) using $\boldsymbol{r}_l$ = *full*, X_{VdW} = 6 Å or $\boldsymbol{r}_l$ = *rot*, X_{cubic} = 10 Å.

region contains a cluster of five water molecules in a tunnel that leads back to a solvent exposed surface of the protein. The monitored regions in the two restraining protocols are broadly similar, with the $\boldsymbol{r}_c$ = *bb* leading to an enlarged monitored region owing to greater fluctuations in the positions of the water molecules. The cyano analogue **6**

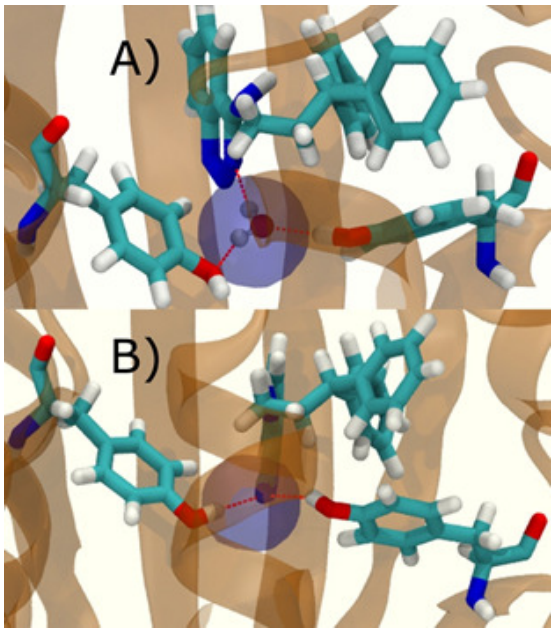


Figure 4.6: Representation of GCT monitored regions in scytalone dehydratase. (A) In complex with **1** (B) in complex with **2**. Regions $\boldsymbol{s}_{AP}$, and $\boldsymbol{s}_{BP}$ are depicted by the transparent blue spheres for $\boldsymbol{r}_c$ = *full* and $X_{medoid} = 4$ Å. The regions obtained with $\boldsymbol{r}_c$ = *bb* and $X_{medoid} = 4$ Å conditions are not shown because they are similar. Relevant hydrogen-bonding interactions between protein residues, water molecules and ligands are depicted by red-dotted lines.

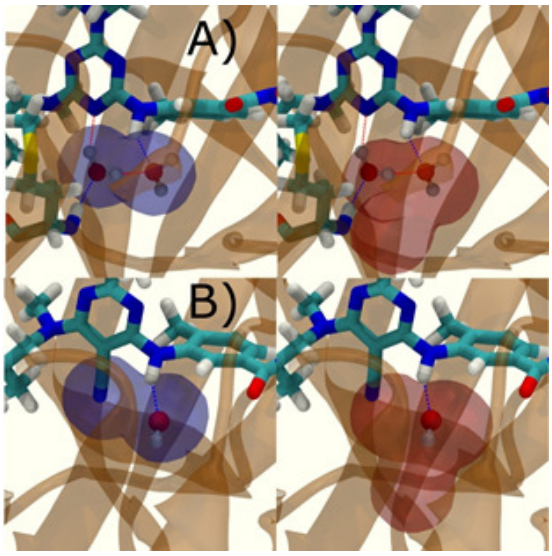


Figure 4.7: Representation of GCT monitored regions in scytalone dehydratase. (A) In complex with **3** (B) in complex with **4**. Regions $\boldsymbol{s}_{AP}$, and $\boldsymbol{s}_{BP}$ are depicted by the transparent blue spheres for $\boldsymbol{r}_c$ = *full* and $X_{medoid} = 4$ Å. The regions obtained with $\boldsymbol{r}_c$ = *bb* and $X_{medoid} = 4$ Å conditions are not shown because they are similar. Relevant hydrogen-bonding interactions between protein residues, water molecules and ligands are depicted by red-dotted lines.

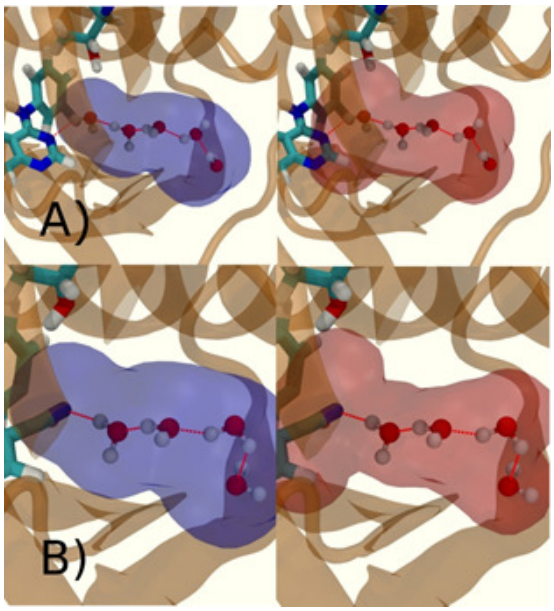


Figure 4.8: Representation of GCT monitored regions in scytalone dehydratase. (A) In complex with **5** (B) in complex with **6**. Regions $\boldsymbol{s}_{AP}$, and $\boldsymbol{s}_{BP}$ are depicted by the transparent blue spheres for $\boldsymbol{r}_c = \textit{full}$ and $X_{medoid} = 4$ Å. The regions obtained with $\boldsymbol{r}_c = \textit{bb}$ and $X_{medoid} = 4$ Å conditions are not shown because they are similar. Relevant hydrogen-bonding interactions between protein residues, water molecules and ligands are depicted by red-dotted lines.

(figure 4.8B) displaces a single water molecule as expected.

4.3.3 Binding energetics

Table 4.1 summarizes the components of the thermodynamic cycle depicted in figure 4.2, for varying restraint protocols and parameters that define the size of the monitored regions. The hydration free energies (rows 1-4) have been discussed previously. These data are completed with protein-ligand interaction energies (rows 5, 6), enabling computation of all the components (rows 7-12) of the thermodynamic cycle depicted in figure 2 for restraint protocols that feature heavy-atom restraints or limited restraints. Comparison of rows 7 and 8 indicate that while interaction energies are broadly consistent for scytalone dehydratase and EGFR kinase with the $\boldsymbol{r}_c = \textit{full}$ or $\boldsymbol{r}_c = \textit{bb}$ protocols, there is a significant variation in the case of p38 MAP kinase. Visualisation of the trajectories indicate that this occurs because **3** adopts a shifted binding mode owing to the occasional decreased water content of the monitored region, and protein side-chain rearrangements. Variations in protein-protein interaction energies are ignored in the present cycle and the result is unbalanced interaction energies between **3** and **4**. The inclusion of protein-protein energies would likely introduce high errors and for this reason longer simulations would be required. Rows 11 and 12 list the resulting binding site water displacement free energies for the three systems with the $\boldsymbol{r}_c = \textit{full}$ or $\boldsymbol{r}_c = \textit{bb}$ protocols. Both protocols indicate that the energetic cost for

Protein	Scytalone dehydratase		p38 MAP kinase		EGFR kinase	
Ligand	**1**	**2**	**3**	**4**	**5**	**6**
$\Delta G^{\boldsymbol{s}_{(A\|B)P}}_{w,(A\|B)P(\boldsymbol{r}_c)}$, $\boldsymbol{r}_c$ = *bb*, X_{medoid} = 4 Å	-3.2 ± 1.2	0	-7.9 ± 1.5	-8.2 ± 0.4	-39.6 ± 2.9	-32.9 ± 3.1
$\Delta G^{\boldsymbol{s}_{(A\|B)P}}_{w,(A\|B)P(\boldsymbol{r}_c)}$, $\boldsymbol{r}_c$ = *full*, X_{medoid} = 4 Å	-4.2 ± 0.2	0	-12.70 ± 0.02	-6.2 ± 0.1	-36.4 ± 0.4	-27.9 ± 1.8
$\Delta G^{\boldsymbol{s}_{(A\|B)}}_{w,(A\|B)(\boldsymbol{r}_l)}$, $\boldsymbol{r}_l$ = *full*, X_{vdw} = 6 Å	-21.2 ± 0.8	-19.6 ± 0.8	-46.6 ± 0.5	-43.9 ± 0.6	-11.8 ± 0.2	-19.1 ± 0.4
$\Delta G^{\boldsymbol{s}_{(A\|B)}}_{w,(A\|B)(\boldsymbol{r}_l)}$, $\boldsymbol{r}_l$ = *rot*, X_{cubic} = 9 Å	-17.7 ± 0.8	-19.3 ± 1.5	-43.5 ± 1.3	-39.4 ± 0.8	-17.4 ± 1.5	-17.9 ± 0.6
$\Delta E(AP \rightarrow BP, \boldsymbol{r}_c)$, $\boldsymbol{r}_c$ = *bb*	-59.9 ± 0.2	-72.3 ± 0.2	-79.7 ± 0.2	-82.2 ± 3.1	-57.9 ± 0.3	-65.8 ± 0.4
$\Delta E(AP \rightarrow BP, \boldsymbol{r}_c)$, $\boldsymbol{r}_c$ = *full*	-61.4 ± 0.1	-75.0 ± 0.1	-68.4 ± 0.3	-83.7 ± 0.2	-54.7 ± 0.1	-62.0 ± 0.1
$\Delta\Delta E(AP \rightarrow BP, \boldsymbol{r}_c)$, $\boldsymbol{r}_c$ = *bb*	-12.39 ± 0.04		-2.5 ± 3.0		-8.0 ± 0.5	
$\Delta\Delta E(AP \rightarrow BP, \boldsymbol{r}_c)$, $\boldsymbol{r}_c$ = *full*	-13.6 ± 0.2		-15.3 ± 0.2		-7.2 ± 0.1	
$\Delta\Delta G_{hyd}(A \rightarrow B, \boldsymbol{s}_A, \boldsymbol{s}_B, \boldsymbol{r}_l)$, $\boldsymbol{r}_l$ = *full*, X_{vdw} = 6 Å	-1.6 ± 1.7		-2.7 ± 0.2		7.3 ± 0.7	
$\Delta\Delta G_{hyd}(A \rightarrow B, \boldsymbol{s}_A, \boldsymbol{s}_B, \boldsymbol{r}_l)$, $\boldsymbol{r}_l$ = rot, X_{cubic} = 9 Å	1.6 ± 1.5		-4.1 ± 1.2		0.5 ± 1.5	
$\Delta\Delta G_{hyd}(AP \rightarrow BP, \boldsymbol{s}_{BP}, \boldsymbol{s}_{AP}, \boldsymbol{r}_c)$, $\boldsymbol{r}_c$ = *full*, X_{medoid} = 4 Å	4.4 ± 0.1		6.48 ± 0.04		8.5 ± 1.7	
$\Delta\Delta G_{hyd}(AP \rightarrow BP, \boldsymbol{s}_{BP}, \boldsymbol{s}_{AP}, \boldsymbol{r}_c)$, $\boldsymbol{r}_c$ = *bb*, X_{medoid} = 4 Å	3.2 ± 1.2		-0.3 ± 1.4		6.7 ± 1.7	
$\Delta\Delta G_b(A \rightarrow B, \boldsymbol{s}_{BP}, \boldsymbol{s}_{AP}, \boldsymbol{s}_A, \boldsymbol{s}_B, \boldsymbol{r}_c, \boldsymbol{r}_l)$, $\boldsymbol{r}_c$ = *full*, X_{medoid} = 4 Å, $\boldsymbol{r}_l$ = *full*, X_{vdw} = 6 Å	-10.8 ± 1.7		-11.6 ± 0.4		8.6 ± 2.2	
$\Delta\Delta G_b(A \rightarrow B, \boldsymbol{s}_{BP}, \boldsymbol{s}_{AP}, \boldsymbol{s}_A, \boldsymbol{s}_B, \boldsymbol{r}_c, \boldsymbol{r}_l)$, $\boldsymbol{r}_c$ = *bb*, X_{medoid} = 4 Å, $\boldsymbol{r}_l$ = *rot*, X_{cubic} = 9 Å	-8.2 ± 2.9		-6.8 ± 2.6		-0.8 ± 2.2	
$\Delta\Delta G_b$, Experimental	-2.0		-2.5		0.6	
$\Delta\Delta G_{\boldsymbol{b}}$, MC/FEP study (OPLS-AA/TIP4P)	-1.2 ± 0.2		-3.0 ± 0.3		1.4 ± 0.2	
$\Delta\Delta G_{\boldsymbol{b}}$, MD/FEP study (ff12SB/TIP4P-Ew)	-3.2		-		-	

Table 4.1: Components of the thermodynamic cycle for evaluation of relative free energies of binding with the GCT approach. All figures are in kcal mol^{-1} and are quoted with one standard error of the mean. Data for MC/FEP relative free energy study comes from the following reference [124].

removing the water displaced by **6** in EGFR kinase is higher than for the displaced water molecules in scytalone dehydratase and p38 MAP kinase. However the free energy cost for displacing a water molecule from p38 MAP kinase is strongly influenced by restraints. This is because, as noted previously, in the $\boldsymbol{r}_c = bb$ protocol the water content of the monitored region exchanges slowly between states with one or two water molecules. Thus on average **4** displaces less than one water molecule under these conditions. The data in rows 11 and 12 can be compared with MC/FEP results from Michel et al. [124], that reported MC/FEP water displacement free energies of 5.5 ± 0.2 kcal mol^{-1} (scytalone dehydratase), 4.2 ± 0.1 kcal mol^{-1} (p38 MAP kinase) and 6.9 ± 0.1 kcal mol^{-1} (EGFR kinase). Quantitative agreement is not expected as the methods used differ, but qualitatively these figures are in closer agreement with those produced by the $\boldsymbol{r}_c = full$ protocol. Completing the cycle yields relative binding free energies (rows 13 and 14). The binding free energies have lower standard errors of the mean in the *full* restraints protocol for scytalone dehydratase and p38 MAP kinase, but not EGFR kinase, presumably because of the larger number of water molecules in the monitored region of the latter protein. The variations of the computed relative binding energies are much greater than observed experimental data (row 15) or those obtained by previous MC/FEP calculations (row 16, albeit with a different forcefield) [124]. An MD/FEP calculation was also run by Stefano Bosisio with identical force fields. However, for the calculation the water had to be contained in the area of interest to calculate the free energy of the water in the site. This caused a discrepancy in both the MD/FEP (row 17) and the MC/FEP with both not matching experiment well due to the effect of restraining the water molecule within the water displacement site. The GCT computed hydration free energies of small organic molecules have been shown previously to be highly correlated to TI computed hydration free energies. This suggests that the discrepancy here is likely due to the neglect of additional contributions such as changes in intramolecular energetics, or protein-ligand entropies, that would normally be included in a FEP/TI calculation. Others have also reported that the use of restraints tends to exaggerate the magnitude of the binding free energies of probe molecules to protein regions, largely due to better probe (solvent) accessibility in sites which are maintained by restraints [128]. Nevertheless, the qualitative picture does not change, and the relative binding free energy for **5** → **6** is much less favourable than for **1** → **2** and **3** → **4**.

4.3.4 Entropic and enthalpic contributions to the energetics of binding site water displacement

The free energy change for water displacement was decomposed into enthalpic and entropic contribution. Figure 4.9A indicates that in almost all cases the enthalpic component is unfavourable, whereas the entropic component is favourable regardless of

the restraining protocol. The only exception is for **3** $\rightarrow$ **4** and $\boldsymbol{r}_c = bb$, where the results are difficult to interpret since the number of water molecules displaced is on average less than one. The entropic component is relatively small and varies little across all systems, and variations in enthalpy changes dominate the overall thermodynamic signature. Figure 4.9B breaks down further the entropy changes into vibrational, librational and orientational components. The results indicate that displacing a water molecule may increase or decrease the vibrational and librational water entropy depending on the binding site and the simulation protocol, but the orientational entropy component dominates the overall entropy variations. This indicates that the favourable entropic contribution upon water displacement is due to the increased number of hydrogen-bonding orientations available to water in bulk.

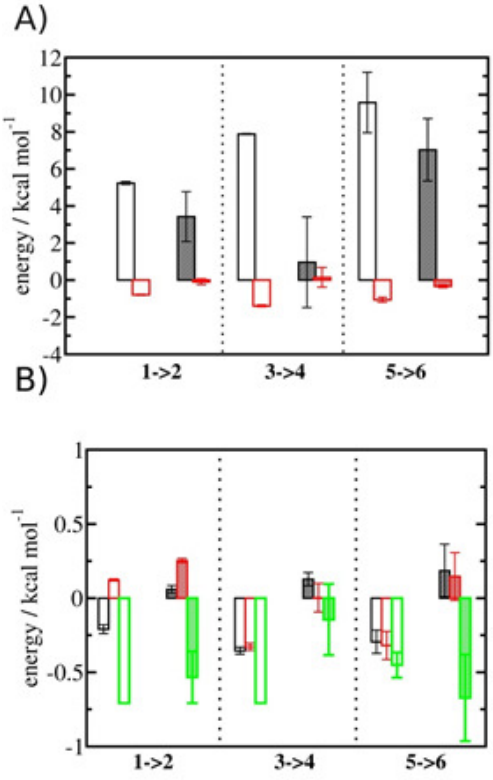


Figure 4.9: Thermodynamic signature of the changes in the hydration energetics of the three protein-ligand complexes. (A) Enthalpy changes $\Delta\Delta H_{hyd}$(AP $\rightarrow$ BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$), are shown as empty ($\boldsymbol{r}_c = full$) or shaded ($\boldsymbol{r}_c = bb$) black histograms. Entropy changes -$T\Delta\Delta S_{hyd}$(AP $\rightarrow$ BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$), are shown as empty ($\boldsymbol{r}_c = full$) or shaded ($\boldsymbol{r}_c = bb$) red histograms. (B) Decomposition of the entropy changes in vibrational entropy -$T\Delta\Delta S_{vib}$(AP $\rightarrow$ BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$) (black), librational entropy -$T\Delta\Delta S_{lib}$(AP $\rightarrow$ BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$) (red) and orientational entropy -$T\Delta\Delta S_{ori}$(AP $\rightarrow$ BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$) (green) components. Error bars represent the standard error of the mean from three replicates.

4.3.5 Localisation of perturbations in water energetics

Additional insights into the binding process are gained by spatial decomposition of the hydration energetics of the monitored regions $\boldsymbol{s}_c$ into sub-regions. Figure 4.10A shows that for p38 MAP kinase, the largest contribution arises from the volume of space $\boldsymbol{s}_0$ (blue) that was occupied by the water molecule displaced by **4**. The cyano group additionally perturbs the interactions of the neighbouring water molecule, shifting it from region $\boldsymbol{s}_1$ (red) towards $\boldsymbol{s}_2$ (green). The net effect almost cancels out and the energetic contributions from $\boldsymbol{s}_0$ are very similar to the *full* monitored region $\boldsymbol{s}_c$. In EGFR kinase (figure 10B) the water volume displaced by the cyano group of **6** (blue) also accounts for the majority of the changes in hydration energetics. Additionally, the first hydration (red) and second (purple) hydration shells of the cyano group are destabilized, whereas the third (maroon) hydration shell is stabilised, and the fourth hydration shell (green) is unperturbed. Thus introduction of the cyano group has perturbed water properties up to 10 Å away. Here water network perturbations (all regions $\boldsymbol{s}_i$, i <0) contribute approximately 1 additional kcal mol^{-1} to the changes in hydration energetics. Thus, that the **5** $\rightarrow$ **6** substitution is not energetically favourable is the result of: higher water displacement energetics (Figure 4.10A $\boldsymbol{s}_0$ versus Figure 10B $\boldsymbol{s}_0$), water network rearrangement penalties (figure 4.10B $\boldsymbol{s}_1$, $\boldsymbol{s}_2$, $\boldsymbol{s}_3$), and weaker improvements in protein-ligand interaction energies (Table 4.1, row 8).

4.4 Discussion

The present study analysed in details the consequences of the use of different restraint protocols to control the allowed flexibility of protein and ligand molecules over the course of an MD simulation. Restraints are undesirable in the sense that they are artificial, and as the results have shown, can quantitatively and qualitatively affect the outcome of a GCT analysis. On the other hand, limited or lack of restraints, that should give more accurate results, leads actually to poor reproducibility of computed quantities for simulations on a ca. 10 ns timescale due to prohibitive amounts of sampling. An important consideration of the present study was to explore the feasibility of using GCT for routine analyses in the context of structure-based drug design programs where computation is typically required to inform the evaluation of hundreds of candidate compounds on a timescale of a few days. In this context, very long MD simulations are not practical. Overall the results suggest that for thermodynamic-cycle analyses, restraints should be used to probe specific protein- ligand conformational states. If different binding modes are to be evaluated, this is best done by separate analyses of different conformational states with $\boldsymbol{r}_c$ = *full* restraints. Alternatively, prohibitively, long simulations may be needed to average over multiple binding modes, as evidenced

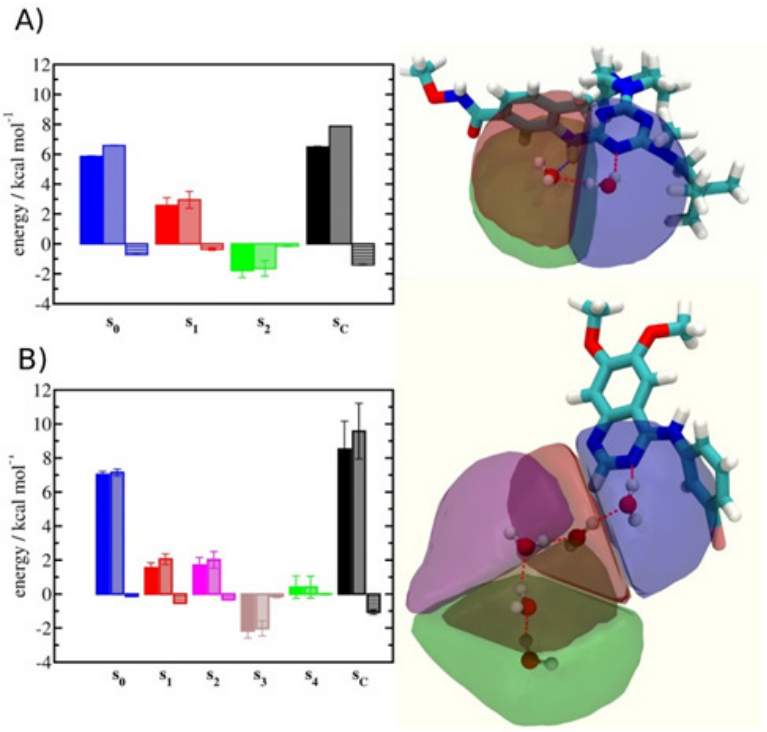


Figure 4.10: Spatial decomposition of the changes in hydration energetics within the GCT monitored regions. The monitored region depicted in figure 4.7 and 4.8 was broken down into sub-regions for the $\boldsymbol{r}_c$ = *full* protocol simulations. For each sub-region, the relative hydration free energy $\Delta\Delta G_{hyd}$(AP $\rightarrow$ BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$), relative hydration enthalpy $\Delta\Delta H_{hyd}$(AP $\rightarrow$ BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$) and relative hydration entropy -$T\Delta\Delta S_{ori}$(AP $\rightarrow$ BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$) are depicted as bars (left). The contributions from the *full* region $\boldsymbol{s}_c$ are shown in black. The right panel depicts the localisation of each sub-region. (A) p38 MAP kinase. (B) EGFR kinase.

for **3** with the protocol that enabled side-chain and ligand flexibility in p38 MAP kinase. If the expected binding modes are unknown, they could be explored prior analyses by means of unrestrained MD simulations. Additionally, care should be taken when selecting a representative conformation of the ligand for solution calculations, as evidenced by the discrepancy in computed relative hydration free energies for **5** and **6**.

Arguably, the appeal of GCT is in the additional information that it provides over, for instance, an alchemical relative hydration free energy calculation. The breakdown of hydration free energies into enthalpic and entropic components revealed that the variations in hydration energetics upon water displacement are dominated by enthalpy. As well the breakdown of the free energy of hydration, the system can be spatially resolved, which may be invaluable for an experience drug designer in the design ofF new ligands.

A rationale for displacing water molecules from binding sites is the associated gain in entropy that should favour the process. However the data shown in Figure 4.9 show that this outcome, at least for the cases investigated here, may only be achieved if the relatively larger loss of enthalpy is counter-balanced by equally favourable ad-

ditional protein-ligand interaction energies. In essence, harnessing entropy by water displacement requires carefully maintaining an energetically similar pattern of hydrogen bonding interactions at the site of the displaced water molecule. The entropy gains are dominated by a favourable increase in orientational entropy and this is due to the lower average number of orientations that a water molecule may adopt in a binding site versus bulk conditions. Such observations have been reported for water in other binding sites, [4], [31] and for a range of idealised host-guest cavities [4]. While it is possible to evaluate enthalpic and entropic contributions to free energies of binding of water molecules with FEP/TI this would require many more simulations at multiple temperatures [129], and this route does not provide a breakdown of entropic contributions into physically insightful translational, rotational and orientational motions.

An important additional insight into the physical chemistry principles that underpin water-mediated protein-ligand interactions is provided by Figure 4.10. In both p38 MAP kinase and EGFR kinase, most of the change in hydration free energy due to water displacement comes from the water molecule that was displaced by the cyano group of **4** and **6** respectively. However, further analysis of the neighboring solvent regions reveal that large but compensating variations in water energetics occurred. In the case of EGFR kinase, the perturbations in water properties propagate up to the third hydration shell of the cyano moiety, and these water network perturbations account for an additional penalty to the water displacement cost of approximately 1 kcal mol^{-1}. Investigation of other systems is desirable to establish the magnitude and frequency of water network perturbation effects in protein-ligand complexes.

4.5 Conclusions

The GCT methodology was developed to provide insights into the hydration thermodynamics of organic and biomolecules. Here it was applied for the first time to a set of protein-ligand complexes where congeneric ligand pairs displace a single water molecule from the binding site. It was shown that protocols that restrain the range of allowed motions of the protein and ligand may be the more judicious choice in cases where throughput and speed considerations are important, as they are for applications to structure-based drug design programs. More realistic models (i.e., fewer or no restraints) will require significantly longer simulations to achieve reasonable reproducibility. While hydration free energies can be predicted with a range of methodologies, the appeal of the GCT technique is that it provides insights into the contributions of enthalpy and entropy to the free energy changes, and that it enables a spatial decomposition of these components. This was used here to determine the spatial extent of the energetics perturbations in a water network upon modification of the chemical structure of a ligand. Further developments of the GCT formalism would be desirable

to account for associated changes in protein and ligand entropy [130], and to automatically assess the conformational dependence of hydration free energies. Overall the current GCT implementation appears well suited for clarifying the role of water in protein-ligand binding, and applications in combination with, for instance, alchemical free energy methods [131] should be envisioned. If GCT and alchemical free energy calculations are combined the solvent contribution to the binding event can be elucidated allowing identification of contributions of solvent, protein, and ligand which are often hard to untangle if an alchemical free energy calculation was run.

Chapter 5

Scoring congeneric ligand-protein series

"...because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know... it is the latter category that tend to be the difficult ones" - Donald Rumsfeld

5.1 Introduction

In this chapter various models of protein-ligand binding event are created using terms derived from GCT (described in chapter 2). The goal of the work here is to find a robust protocol capable of prediction (ranking of ligand binding affinities) *a priori* in either a ligand optimisation or ligand discovery context. The types of simulations, and the kind of restraints required to give a reproducible and accurate protocol are the prominent questions investigated. Restraints tend to exaggerate interaction energies [115] and affect the accuracy of the method. On the other hand lack of restraints necessitates long simulation times to achieve acceptable precision. Here a search for the optimum balance between precision and accuracy of a GCT model of protein-ligand binding is sought. The transferability of protocols between different protein-ligand systems is also a key issue investigated.

To do this, two congeneric series (ligands of the same scaffold), factor Xa (FXa) and heat shock protein 90 (HSP90) series, of ligands are studied. First a series of FXa ligands were analysed because of historical data available from studies by Abel [132], and Nguyen [53] respectively using the IFST method to predict hydration thermodynamics.

However, both of these papers, use extensive machine learning to optimise either the grid cutoff [53] for the enthalpy or entropy, or otherwise the selection of parameters in multiparameter models [132]. Here this approach was avoided because it limits the applicability of the methodology since training sets are required.

Another issue is practicality in industry. Use of computation in an industrial context is limited by time constraints, essentially the speed in which a compound can be generated and tested. This is about a week in a typical ligand optimisation context. This implies that with current technology available to a pharmaceutical company, MD simulations up to the scale of 10-100ns/ligand seem practical assuming accurate prediction of improved binders (70% correct ranking predicted).

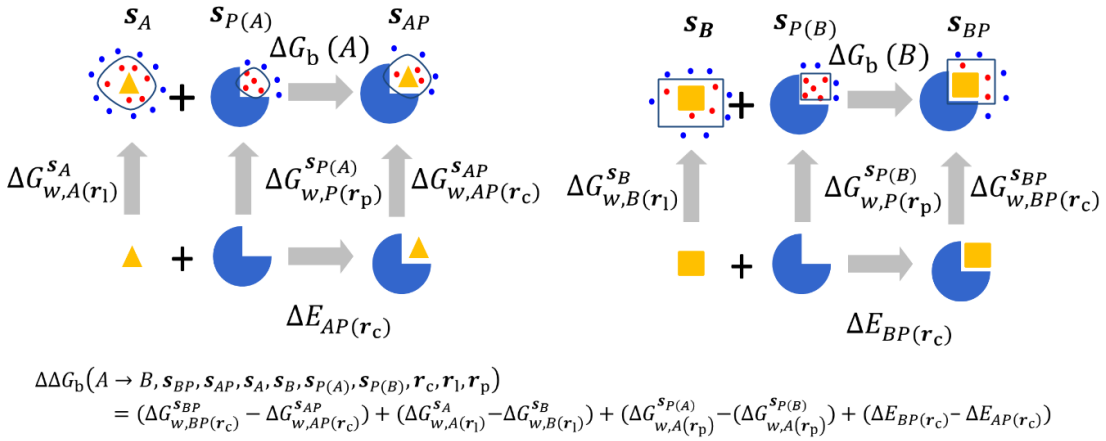
To test transferability, another congeneric series was considered. This series is composed of inhibitors of HSP90a. This system was interesting because several stable waters involved in bridging interactions between the ligand and protein were found in several crystal structures in the series. Also the series is important because HSP90a is involved in many cancers and is a very well known drug target of interest to the pharmaceutical industry. Brief introductions to the pharmaceutical relevance of each system are provided.

5.1.1 Factor Xa, a coagulation factor

FXa is an important molecule in the thrombosis pathway. It is especially important in cases of thrombosis where blood clots are formed, restricting blood flow. FXa activates prothrombin which reduces thrombin generation reducing thrombus growth i.e, clot formation [133]. Direct FXa inhibitors are good anticoagulants and for this reason it is a good therapeutic target which could be used to treat various thrombosis related diseases.

5.1.2 Heat Shock Protein 90a

HSP90a is a vital chaperone protein in the cell which regulates many cellular pathways. HSP90a folds many upregulated oncogenic proteins in cancers so its inhibition could provide a useful therapeutic in many cancers. HSP90a contains a N-terminal ATPase binding site with several buried waters of various stabilities. These buried waters could play important roles during the ligand binding event. Several ligands of a HSP90 series were investigated [19, 134]. Simple ligand mutations such as changing an oxygen to an amine group (which interacts with a buried water) change the binding affinity drastically suggesting a significant contribution from differences in binding-site water energetics.



$$\Delta\Delta G_b(A \to B, s_{BP}, s_{AP}, s_A, s_B, s_{P(A)}, s_{P(B)}, r_c, r_l, r_p)$$
$$= (\Delta G^{s_{BP}}_{w,BP(r_c)} - \Delta G^{s_{AP}}_{w,AP(r_c)}) + (\Delta G^{s_A}_{w,A(r_l)} - \Delta G^{s_B}_{w,B(r_l)}) + (\Delta G^{s_{P(A)}}_{w,A(r_p)} - (\Delta G^{s_{P(B)}}_{w,A(r_p)}) + (\Delta E_{BP(r_c)} - \Delta E_{AP(r_c)})$$

Figure 5.1: Thermodynamic cycles for evaluation of relative terms between two ligands A and B: relative ligand desolvation free energies, relative protein desolvation free energies, relative complex solvation energies, relative interaction energies, and relative free energies of binding. Ligands are depicted by yellow shapes. Proteins are depicted by blue shapes. In all GCT analyses, water molecules (red circles) inside the monitored regions, $s_{BP}, s_{AP}, s_A, s_B, s_{P(A)}, s_{P(B)}$ contribute to the computed hydration free energies, whereas those that are out of the monitored regions in blue are ignored. Different restraint protocols (r) may be used to control allowed protein (p), ligand (l), and complex (c) motions.

5.1.3 Theory

In this chapter similar relative (ligand B - ligand A) binding thermodynamic cycles are used as demonstrated in chapter 4. The only major difference is the addition of a relative protein desolvation term as shown in Figure 5.1. This thermodynamic cycle again assumes rigid ligands and proteins and contains four relative thermodynamic terms which add up to a relative binding free energy. These are the relative ligand desolvation free energies ($\Delta G^{s_A}_{w,A(r_l)} - \Delta G^{s_B}_{w,B(r_l)}$), relative protein desolvation free energies ($\Delta G^{s_{P(A)}}_{w,P(r_p)} - \Delta G^{s_{P(B)}}_{w,P(r_p)}$), relative complex solvation energies ($\Delta G^{s_{BP}}_{w,BP(r_c)} - \Delta G^{s_{AB}}_{w,AP(r_c)}$), and relative interaction energies ($\Delta E_{BP(r_c)} - \Delta E_{AP(r_c)}$), which can be combined to give the relative free energies of binding $[\Delta\Delta G_b(A \to B, s_{BP}, s_{AP}, s_A, s_B, s_{P(A)}, s_{P(B)}, r_c, r_l, r_p]$.

Different combinations of these relative terms are tested to see if a good predictive value can be obtained. Next, the protocol for the preparation of the molecular models is presented.

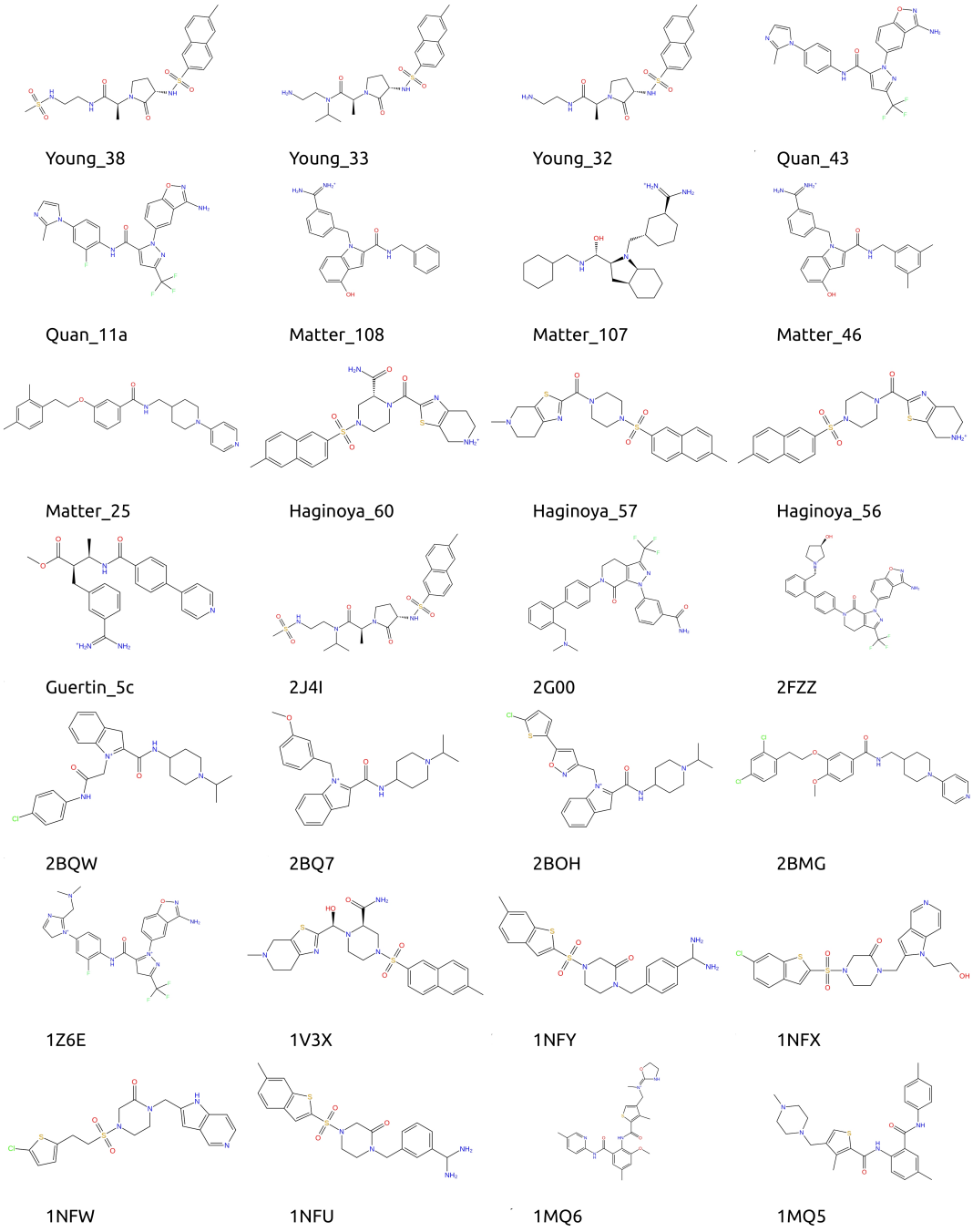
5.1.4 Preparation of FXa simulations

5.1.4.1 Preparing the FXa "pseudo apo" (PSAPO) structure

The PSAPO, "pseudo apo", simulation used the same protein structure, 1FJS [20], as that found in two previous studies [53, 132] as an initial input. A PSAPO structure is essentially a protein structure obtained after a ligand is removed from a complex. In this way a series of congeneric ligands can be tested for the overlap of protein waters which must be removed. However, this assumes that the protein relaxes to a similar configuration (rigid body assumption). The 1FJS structure was used for PSAPO (protein alone) simulation with the ligand removed. The structure without the ligand was parameterised using the software 'tleap' to have Amber 11 [84] protein parameters using the parameters from AMBER99SB forcefield [135]. This system was solvated with TIP4P-EW [81] waters in a rectangular box with the edges of the box extended at least 11 Å away from the edges of the protein. Four disulfide bonds (covalent bonding of sulfur atoms) were generated to be consistent with the crystal structure. These involved the following cysteine residues: 7,12; 27,43; 156,170; 181,209. No ions were incorporated in the simulations, even for charged systems. The system was first energy minimised with AMBER using FESetup [127].

5.1.4.2 Preparing FXa ligand simulations

FESetup was also utilised to generate ligand input files of the same ligand pairs described by Nguyen *et al.* [53] as shown in Figure 5.2, using the GAFF forcefield [44] and AM1-BCC charges [45], as implemented in the AMBER11 software suite [84]. In contrast to previous papers all ligand and complex simulations were done using alternative protonation states for the following ligands: 1MQ5, 1MQ6, 1NFU, 1NFX, 1NFY, 1V3X, 1Z6E, 2BMG, 2BOH, 2BQ7, 2BQW, 2GOO, Haginoya_57, and Matter_25. This was due to unclear protonation states which were experimentally ambiguous as well as difficult to interpret with pKa predictions [136] where often a ratio of protonation states was expected at a pH of 7 (simulations with protonation states equivalent to studies by Nguyen *et al.* [53] and Abel *et al.* [132] were run but correlations were poor so are not discussed). Also, it is known that concentrations of protons would require simulation boxes which are not computationally tractable making protonation states difficult to predict using simulation. The ligand conformation found in the minimised complex was used for the ligand simulations because it is more consistent with the rigid thermodynamic cycle described in figure 1.2 in chapter 1. The ligands were solvated in a rectangular box with TIP4P-EW waters. The edges of the box extended at least 11 Å away from the edges of the ligand, identical to the procedure used in the PSAPO preparation.

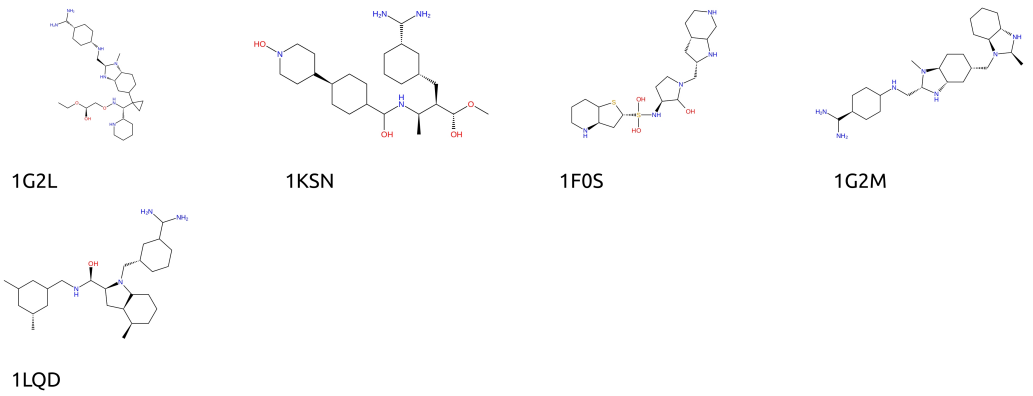


(a)

Figure 5.2: 2D structures of all ligands which match the FXa dataset presented by Abel *et al.* [132] and Nguyen *et al.* [53] as generated from Maestro software created by Schrödinger LLC. However, the protonation states of the following ligands differ from those of Abel *et al.* [132] and Nguyen *et al.* [53]: 1MQ5, 1MQ6, 1NFU, 1NFX, 1NFY, 1V3X, 1Z6E, 2BMG, 2BOH, 2BQ7, 2BQW, 2GOO, Haginoya_57, and Matter_25.

5.1.4.3 Preparing FXa complex simulations

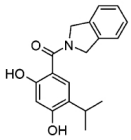
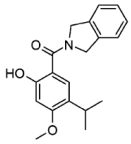
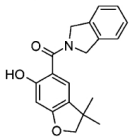
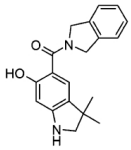
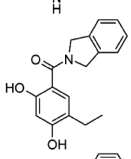
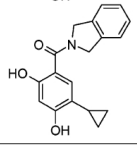
FESetup was also used to automatically setup the complex simulations for all FXa ligand pairs. Again a rectangular box of TIP4P-EW waters was generated in the



(b)

Figure 5.2: continued structures for FXa ligands

same way as for the FXa ligand simulations. Minimisation with steepest descents and conjugate gradient used with the general AMBER 11 implementation so that there are no clashes between the protein, ligand and waters in the system.

Compound	Structure	Hsp90[b] K_d (μM)	HCT116[a] IC_{50} (μM)
1		0.00054	0.031
2		0.015	2.5
3		1.3	0% at 10 μM
4		0.31	25
5		0.002	0.23
6		0.004	0.18

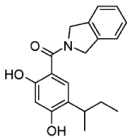
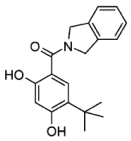
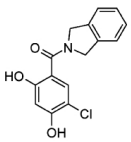
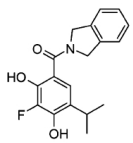
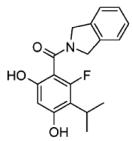
Compound	Structure	Hsp90[b] K_d (μM)	HCT116[a] IC_{50} (μM)
7		0.003	0.08
8		0.001	0.29
9		0.004	3.5
10		~	30
11		~	61% at 100 μM
36[c]	17-DMAG	0.21	0.05

Figure 5.3: All HSP90a ligands in the dataset with cell assay and IC50 experimental data and image adapted from [19].

5.1.5 Preparation of HSP90a simulations

Similar simulation protocols were followed as for Factor Xa with minor differences explained below. All ligands which were simulated are shown in Figure 5.3, with the exception of ligand **7** which had stereoisomers so was removed from the dataset as it was ambiguous which isomer bound to HSP90a and also ligand **36** a ligand of a different scaffold(HSP90a ligand will be referred to by number). For the PSAPO simulation the 2XAB [19] pdb structure was used. All relevant complex structures were generated using the same protein structure as the initial protein structure (PSAPO). Ligands were aligned onto the ligand present in the 2XAB structure. Another difference was that each protein or complex was solvated in a rectangular box whose edges extended 12 Å from the edge of the solute, which is larger than the long-range nonbonded cutoff used in simulation of 10 Å.

5.1.6 Restraint protocols

GCT calculations were performed with several different protocols that vary in their use of restraints to control the conformations sampled by the ligands or protein during the simulations (identical restraints as in section 4.2.2 of chapter 4). GCT calculations can in principle be performed without any restraints on solutes; however this has a number of disadvantages. Firstly, extensive conformational sampling is required to obtain converged water properties for flexible solutes. Secondly, graphical analyses of voxel properties are more complex. Thirdly, the thermodynamic cycle depicted in Figure 5.1 does not include contributions from changes in conformations or flexibility from the protein and ligands. On the other hand restraints are artificial and may negatively affect the predictions of free energies of binding. In the present work different restraining protocols **r** were compared in an effort to identify a practical protocol for routine calculations.

In the $\boldsymbol{r}_{\mathrm{c}} =$ *full* protocol, positional restraints were applied to all heavy atoms of both ligand and protein. In the $\boldsymbol{r}_{\mathrm{c}} =$ *bb* protocol, positional restraints were applied to only the heavy atoms of the backbone protein. In the $\boldsymbol{r}_{\mathrm{l}} =$ *full* protocol, ligands were restrained in their binding site conformation. Restraints were implemented with a force constant of 10 kcal mol^{-1} Å^{-2} for the *bb* and *full* protocols. All restraints were applied on absolute Cartesian coordinates.

5.1.7 Molecular dynamics simulations parameters used in both systems

All molecular simulations were produced using the software Sire/OpenMM which results from linking the general purpose molecular simulation package Sire (revision 1786), with the GPU molecular dynamics library OpenMM (revision 3537) [89]. Simulations were run at a pressure of 1 atm and temperature of 298 K using an atom-based generalized reaction field nonbonded cutoff of 10 Å for the electrostatic interactions [51], and an atom-based nonbonded cutoff of 10 Å for the Lennard-Jones interactions. A velocity-Verlet integrator with a time step of 2 fs was used. Temperature control was achieved with an Andersen thermostat with a coupling constant of 10 ps^{-1} [48]. Pressure control used attempted isotropic box edge scaling Monte Carlo moves every 25 time steps. The OpenMM default error tolerance settings were used to constrain the intramolecular degrees of freedom of water molecules. For each HSP90a system three simulations of 22 ns were run. Only a single 22 ns simulation for the FXa systems was run. Using the same starting conformation but a different random velocity assignment snapshots were stored every 1 ps and were written in a DCD format. The first ns of each trajectory was discarded to ensure the system was well equilibrated prior to sampling for all simulations. For HSP90a systems, standard errors of the mean were obtained from the triplicate simulations which were also done for each restraint condition. For statistical analysis of the FXa system the single trajectory is chunked into thirds (7 ns sampling each) to get appropriate errors also for each restraint condition.

5.2 Scoring methodologies

Multiple methodologies were tested to see whether predictive rankings of binding affinities can be extracted from limited sampling obtained with 22 ns runs as these were judged to be the maximum time per ligand for a practical workflow for industrial application. In all cases different relative hydration free energy terms are used to predict relative binding affinities of ligand pairs. For the FXa system, pairs are identical to those found in Nguyen *et al.* [53]. In the HSP90a case all ligand pairs are formed relative to ligand **1**.

In the FXa simulations an alternate implementation of the relative protein desolvation free energies was also tested because of its good reported reproduciblity and ease of use [132]. However, in general for both systems vdW methods of selecting grid points were used for all (see chapter 2).

5.2.1 Selection of grid regions with vdW protocols

The vdW method was usually used to select grid regions for data analysis and is the same as described in section 2.2 in chapter 2. In all of these methods the ligand's initial minimised conformation is used for the vdW overlap with grid points. All grids are aligned to the PSAPO simulation grid prior to analysis. This protocol was used for both the $\boldsymbol{r}$=*bb* and $\boldsymbol{r}$=*full* restraints protocol.

5.2.2 Watermap methodology (PSAPO-Abel)

The *ab initio* methodology described by Abel et al. [132] was also implemented. Essentially, a distance between a heavy atom of the ligand and a water oxygen of a predicted site is used to estimate the overlap of a water. This site has an associated free energy of hydration cost computed by IFST or in this case GCT. The amount of overlap between the predicted site and ligand heavy atom is quantified to give an esimate of the desolvation cost of binding site water molecules. To determine overlaps first, density clustering is done on a molecular dynamics trajectory to identify waters within the binding site. This clustering method is identical to that found in section 4.2.5. Briefly reviewing the process:

1. Pick grid points with density $> 2\times$ bulk density
2. Assign all neighbouring grid points of a radius 1.5 Å (or other chosen cutoff) to each high density point to create a site.
3. Select next unassigned high density point
4. Terminate when no points with density $> 2\times$ bulk density are left.

These sites are then assessed using GCT to determine the free energies of hydration sites ΔG_{hs}. Distances between ligand heavy atoms and hydration sites are linearly scaled so that the smaller the distance the closer to the entire water displacement cost (the negative of the hydration free energy of the site). However, a contribution from a single hydration can never exceed the maximum displacement cost of the site itself. This is given by eq. 5.1 used by Abel *et al.* [132]:

$$\Delta G_{\text{bind}} = \sum_{lig,hs} \Delta G_{hs} \left(1 - \frac{|\vec{r}_{lig} - \vec{r}_{hs}|}{R_{\text{co}}}\right) \Theta(R_{\text{co}} - |\vec{r}_{lig} - \vec{r}_{hs}|) \tag{5.1}$$

where ΔG_{bind} is the binding free energy of a ligand and R_{co} is the distance cutoff for when a ligand heavy atom starts to displace a hydration site. (Abel *et al.* [132]

rationalised it as about 2.24 Å calculated by noting that the radii of both the oxygen water and carbon atom 1.4 and leaving a tolerance of 0.8, 0.8 x (1.4+1.4) = 2.24 Å.) $\vec{r}_{lig}$ and $\vec{r}_{hs}$ are the position vectors for the ligand atom and hydration site respectively, while Θ is a Heaviside step function which equates to zero for a negative argument and 1 for a positive argument. It is important to observe that ΔG_{hs} is the free energy of water displacement (energy cost for removal), from a 'PSAPO' simulation. So, the values for each hydration site used by GCT should have the opposite sign because hydration free energies are computed in the GCT implementation. Also, this methodology assumes that the desolvation cost of the water site can be linearly fit. How severe this assumption is may be dependent on system but is unclear.

Clustering was done differently from the Abel paper, following the same protocol as that in chapter 4. Various cluster density cutoffs for centroids and a cluster radius of 1 Å were used. By contrast the Abel paper only used a cluster radius of 1 Å where a cluster point must have a density at least two times greater than bulk water density.

5.2.3 Estimation of relative protein desolvation energetics (PSAPO)

Relative protein desolvation free energies were determined using the vdW method of selecting grid regions. The distance criterion, X_{vdW}, from the vdW surface of each ligand, is varied for simulations performed under both $\boldsymbol{r}_{\mathrm{p}} = full$ and $\boldsymbol{r}_{\mathrm{p}} = bb$. An alignment of all FXa ligands in the binding site is shown in Figure 5.4.

5.2.4 Estimation of relative ligand desolvation energetics (LIG)

Relative ligand desolvation free energies for FXa and HSP90a systems were determined using the vdW method and where applied only with the *full* restraint protocol. The complex minimised ligands (Figure 5.4) were used to generate the relative ligand desolvation free energies.

5.2.5 Estimation of relative protein-ligand hydration energetics (HOLO)

Variants of the vdW protocol were tested for estimating the relative protein-ligand hydration energetics. Visual explanation of the differences in each variant is show in Figure 5.5.

Four variants of the vdW HOLO method were utilised.

1. vdW: This method selects grid points within the vdW distance from all ligand atoms.

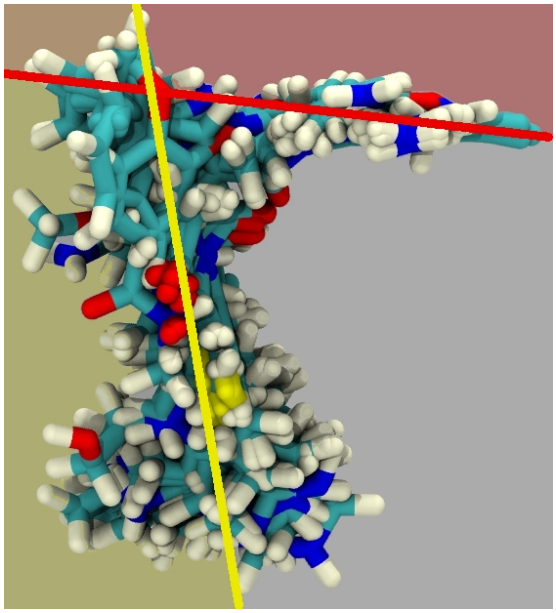


Figure 5.4: Shows the alignment of the complex minimised FXa ligands used to generate PSAPO-Abel, PSAPO, HOLO, LIG energetics. Colours denote planes used in two HOLO variants explained in methodologies. The plane defined in the red is used in the B-Plane HOLO method, while the plane which is solvent exposed is coloured in red, for the S-plane HOLO method. The region where these two planes overlap has been coloured in orange. The area in grey are not solvent accessible because of tight protein-ligand binding.

2. Polar: Same as 1 but only polar atoms are considered.

3. S-plane (solvated plane): Same as 1 but a plane is defined by the alignment of the FXa ligands (Figure 5.4 defined in yellow). All grid points to the right of the figure are no longer considered (since they are solvent inaccessible due to the ligand binding tightly to the protein) so that only solvent exposed areas are considered.

4. B-plane (buried plane): Same as 3 but a different plane is chosen and a red area in Figure 5.4 is only considered. It is referred to B-Plane because only occasional buried waters are generally found in that area.

5.2.6 Estimation of relative protein-ligand energetics (IE)

The interaction energies are extracted from the simulation using Sire. Estimates were performed with both the $\boldsymbol{r} = bb$ and $\boldsymbol{r} = full$ restraint protocols.

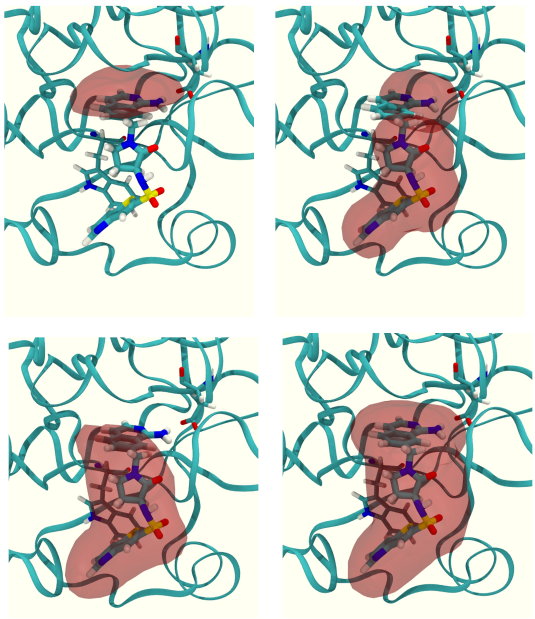


Figure 5.5: Shows the four variants of vdW selection protocol used: on the top left is the B-plane method, top right is the Polar method, bottom right is the normal vdW, and bottom left is the S-plane method.

5.2.7 Combination Analysis

The different relative energetic terms (LIG, PSAPO, HOLO, IE) were combined in different models to predict relative binding free energies. The PSAPO-Abel model was not used in the combination analysis because its predictive value was poor; see the discussion below. For a given combination, the effects of varying the cutoffs (X_{vdW}) were also considered. The predictive power was assessed by evaluating correlations with experimental data. In addition, selected pairs of ligands were analysed in detail to clarify the convergence properties of the energetic terms.

5.2.8 Assessing predictive value

Two methods were used to assess how predictive a model is. First the R^2 value gives an understanding of how well the model correlates with experimental relative binding affinities. This is done by plotting the term or combination against the experimental value, where in this case a linear relationship is expected. Another statistic used is the predictive index, (PI) [137]. This method gives a rank-order statistic which penalises for incorrect ranking and rewards for correct ranking of a pair of ligands and has a value which ranges from -1 to 1, where -1 indicates that all pairs were incorrectly ranked, 0 indicates the ranking was random and 1 indicates a perfect ranking. The predictive

index is calculated following eq 5.2:

$$\mathrm{PI} = \frac{\sum_{j>i} \sum_i w_{ij} C_{ij}}{\sum_{j>i} \sum_i w_{ij}} \tag{5.2}$$

$$w_{ij} = |E(j) - E(i)| \tag{5.3}$$

$$C_{ij} = \ 1 \ if \ \frac{E_j - E_i}{P_j - P_i} < 0 \tag{5.4}$$

$$C_{ij} = -1 \ if \ \frac{E_j - E_i}{P_j - P_i} > 0 \tag{5.5}$$

$$C_{ij} = 0 \ if \ P_j - P_i = 0 \tag{5.6}$$

where the weight is shown in eq 5.3, and E represents experimental results of two compounds i and j. The weights give greater importance to correct rankings of pairs that have significant differences in experimental data. The PI statistic gives more precedence to ranking all ligands in the correct order while R^2 is a statistic which shows how well correlated the predicted values are in comparison to the experimental values.

5.3 Discussion

5.3.1 Factor Xa

The binding poses of Factor Xa ligands are shown in figure 5.4. From the alignments it can be seen that the binding site is not deep and solvent exposed. Typically these ligands bind by adopting a L-shaped pose in the binding site [138].

In this work, 28 ligand pairs are assessed by evaluating whether various combinations of hydration descriptors are predictive. The predictive value of a particular model is assessed with its R^2 and PI values. Since full thermodynamic cycles are not modelled, quantitative predictions are not expected. Instead the aim is for a reasonable ranking of protein-ligand binding affinity which can describe trends in terms of the various types of solvation energetics of the binding process. Each model is systematically assessed for its reproduciblity, and its predictive value.

All combinations of four main descriptors (HOLO, LIG, PSAPO, IE) were used to generate various models annotated in table 5.1, whose symbols will be used throughout to refer to a respective model.

Model	Symbol
PSAPO	a
IE	b
LIG	c
HOLO	d
PSAPO + LIG + IE	e
HOLO + PSAPO	f
HOLO + PSAPO + IE	g
HOLO + IE	h
HOLO + LIG	i
HOLO + LIG + PSAPO + IE	j
HOLO + LIG + IE	k
PSAPO + IE	l
PSAPO + LIG	m
LIG + IE	n

Table 5.1: Table showing symbols which will be used to represent particular models

5.3.1.1 Convergence of hydration energies

Just the single descriptor models (models a-d) were tested for the convergence of the relative hydration energies. First, the PSAPO and HOLO energies were computed as a function of the size of the monitored region defined by the X_{vdW} cutoff parameter and for each restraint protocol. The results are depicted in Figure 5.6. First, the HOLO model is investigated. For several ligands the HOLO term is very noisy for high X_{vdW} cutoffs with uncertainties of $\pm$ 15 kcal mol^{-1}. This is because these cutoffs select grid regions where more bulk waters are included. For, this reason vdW HOLO variant cannot be used in a predictive model. The PSAPO model is much more precise because it uses a single simulation for all ligands while the HOLO model requires two simulations for each relative pair, introducing sampling error. The protein-ligand simulations are also noisier in general because water exchanges around the ligand may involve interactions between the ligand, protein and the water which leads to slower sampling. The PSAPO simulation also benefits from being restrained in a more open "ligand adopted" conformation, which allows quicker water molecule diffusion, helping sampling. The large variation in the magnitude of scores is seen to be dependent on the the restraint protocol. The $\boldsymbol{r} = bb$ conditions generates much more noise in the HOLO energies at any X_{vdW} distance greater than 3 Å because more conformations can be adopted when the side chains of the residues are unrestrained. In general it is difficult to understand with both the APO and HOLO models at which distance from the vdW surface it is appropriate to set a cutoff. However, the data from chapters 3-4 suggest that most of the contribution would come from hydration free energies from the first solvation shells and this suggests that adequate cutoffs typically should range from 0 to 5 Å.

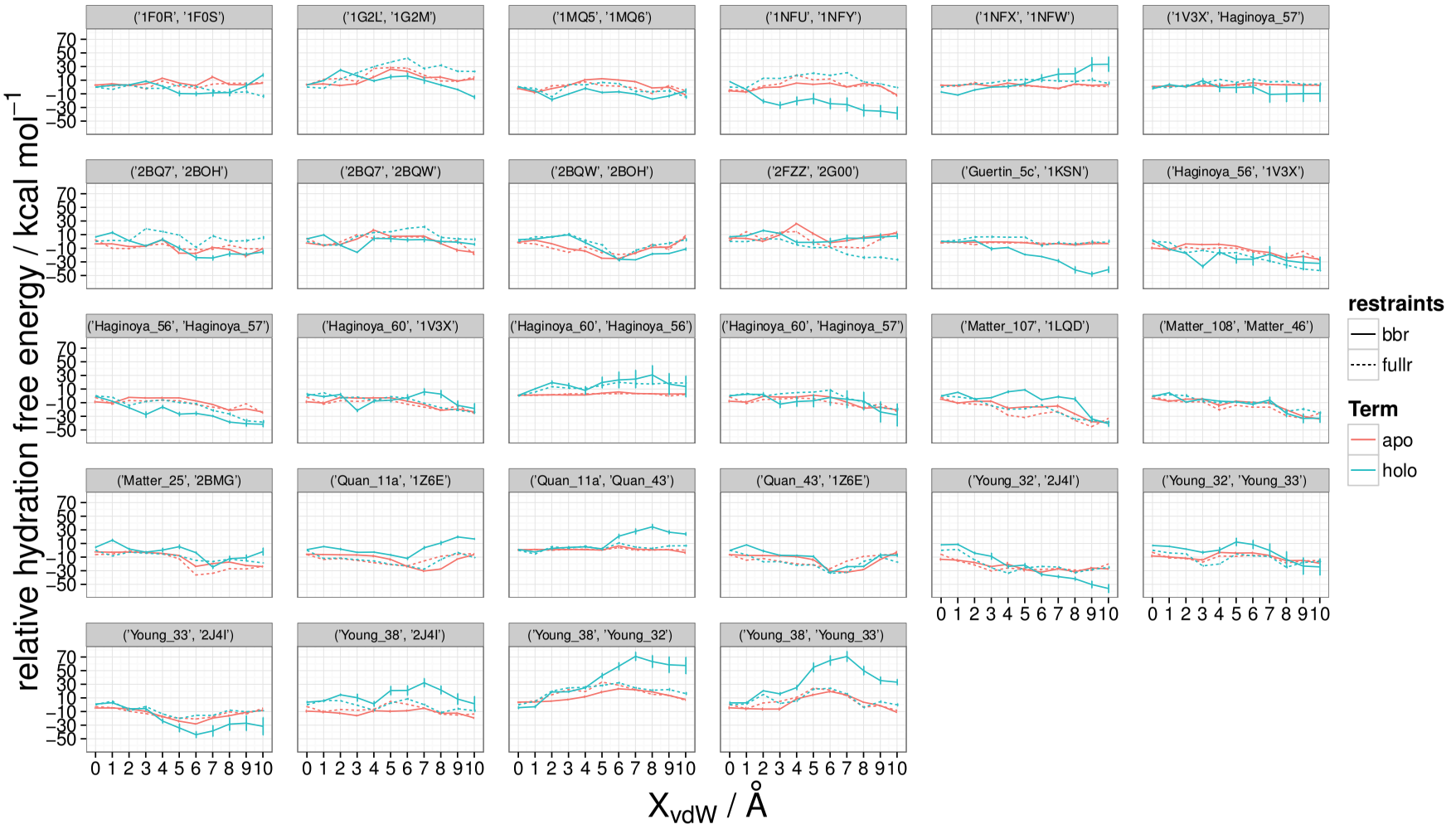


Figure 5.6: The computed APO and HOLO relative hydration free energies as function of the size of the grid region. The HOLO and APO energies are shown as the difference in hydration free energies of ligand 2 minus ligand 1.

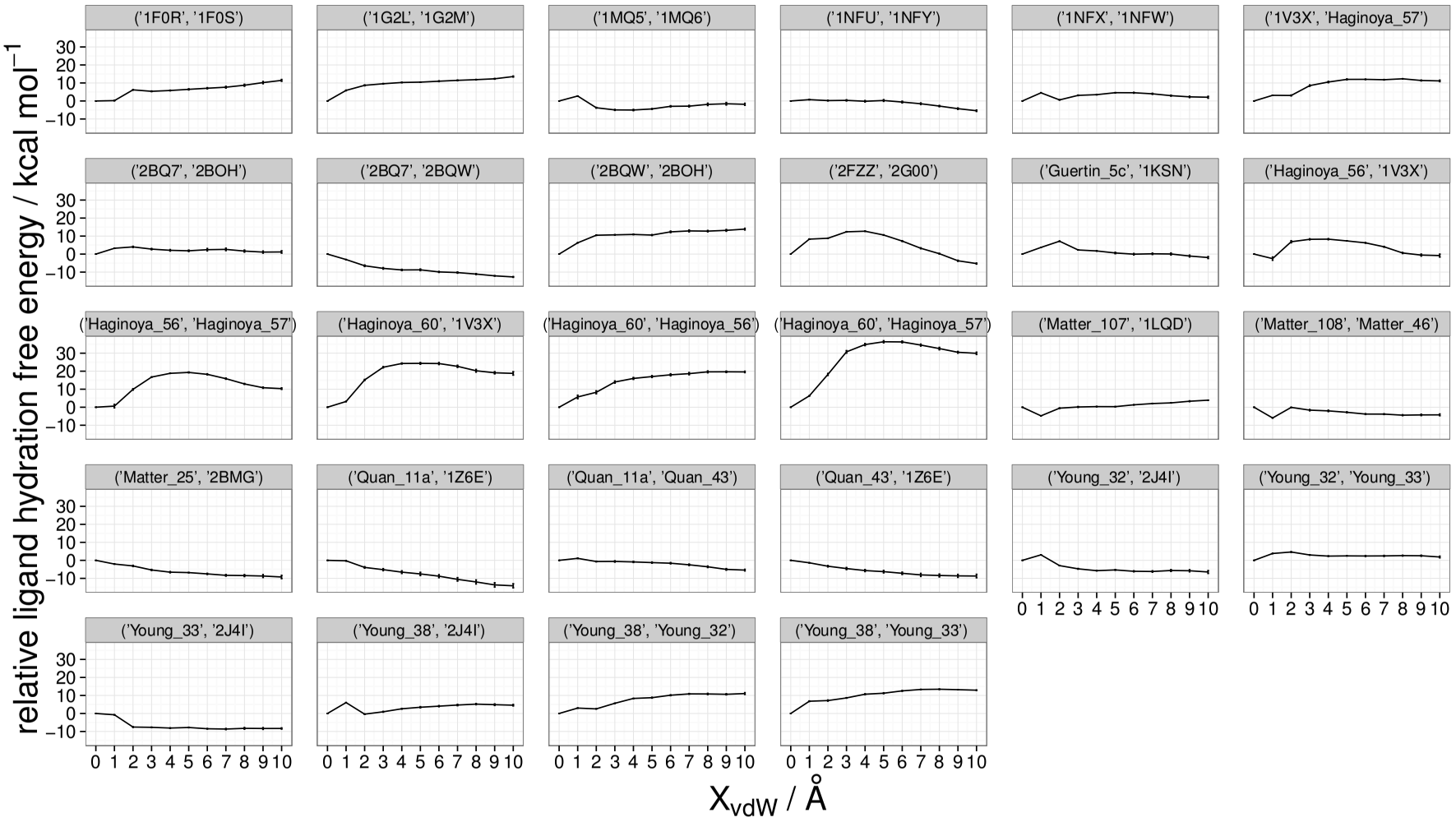


Figure 5.7: The computed LIG relative hydration free energies as a function of the grid region. The LIG energies are shown as the difference in hydration free energies of ligand 2 minus ligand 1.

Next the convergence of the LIG model was analysed as shown in Figure 5.7. The relative ligand desolvation term tends to converge at around 5 Å using the vdW protocol for most ligand pairs. The convergence of the relative of the ligand solvation is shown for every ligand pair. This term had the best convergence for most ligand pairs. There were a few cases (('2FZZ', '2GOO'), ('Quan_11a', '1Z6E')), of some negative drift after 5 Å which is maybe a systematic error in the reference water bulk enthalpy as discussed in chapter 3. However, for lower cutoffs the LIG term is very reproducible. From this data the difficulty in obtaining convergence relates to two things. First of all it shows the limits to the amount of sampling required which means higher computational expense. Secondly, terms which are harder to converge such as the HOLO have to be focused to reduce noise but this may be at the expense of missing important hydration behaviour which is not local. However, terms containing the APO and LIG terms should be easier to reproduce.

5.3.1.2 Evaluation of energetics using PSAPO-Abel model

The PSAPO-Abel model, was tested for comparison with published data from Abel *et al* [132]. Several issues were encountered since there are discrepancies in the protonation states of the ligands described by Abel *et al.* [132]. They suggest several protonated and charged ligands however, it is unclear whether these are true in solution (as mentioned in pg 83). Alternative protonation states were used here, (see Figure 5.2) for all the structures but those protonation states defined by the Abel *et al.* [132] study were also tested.

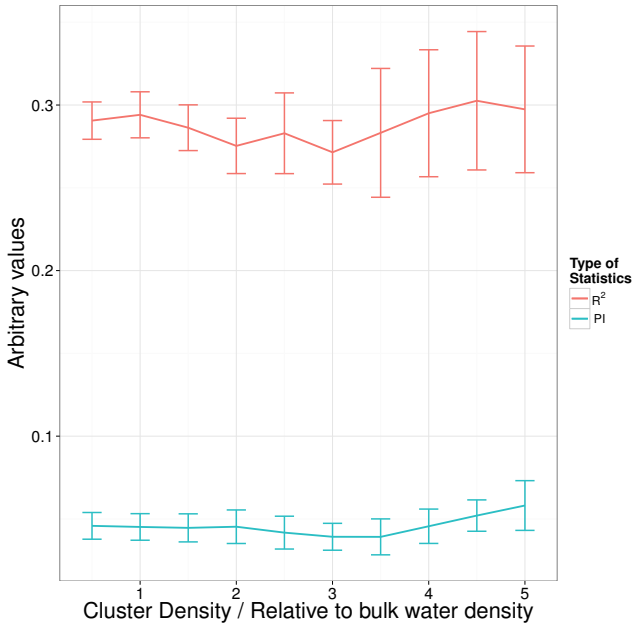


Figure 5.8: Effect of various cluster densities for centroid discovery prior to watermap apo calculation

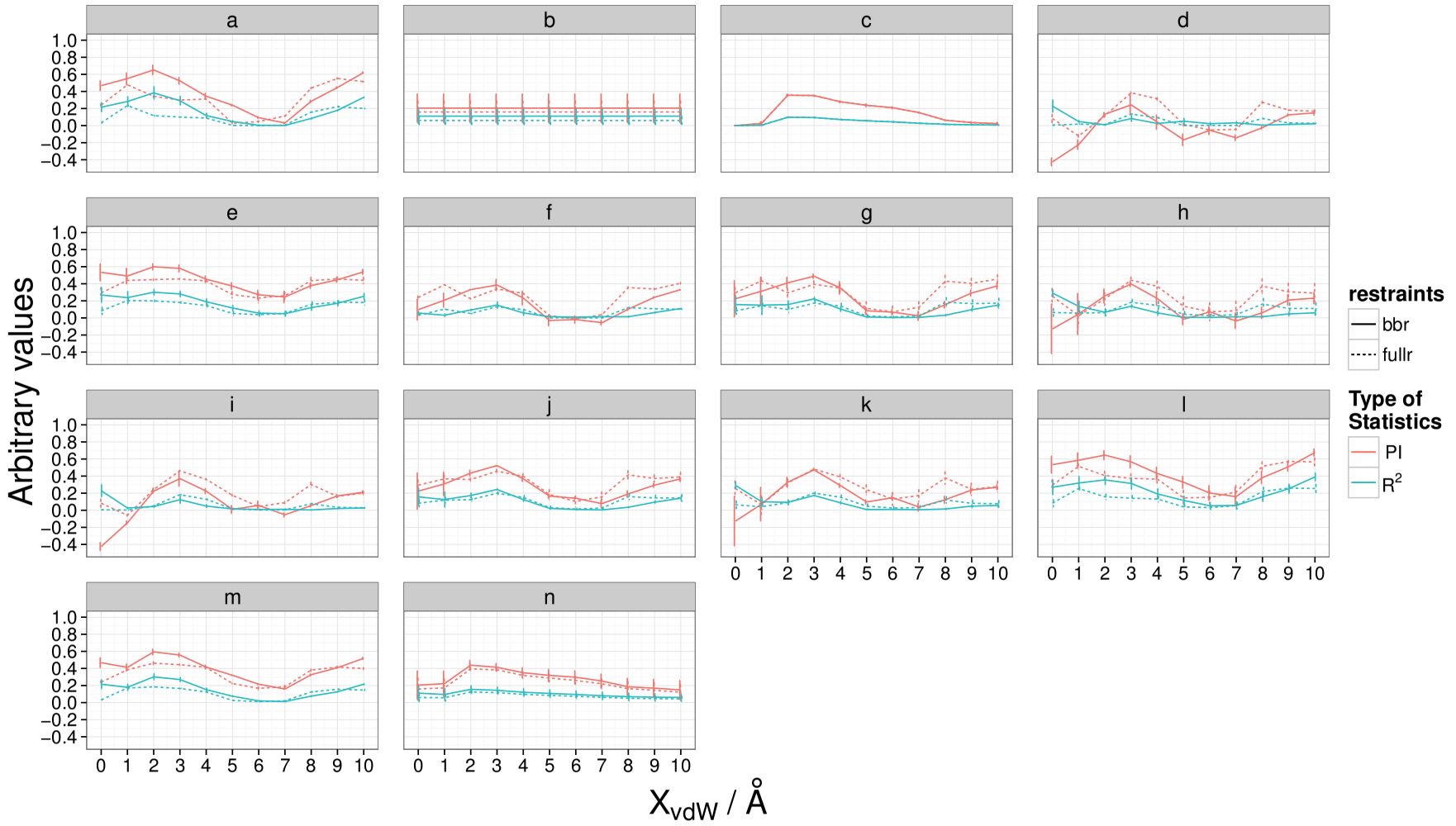


Figure 5.9: The predictive power of all 14 models as a function of the size of the monitored grid region. Labels represent model as shown in table 5.1. Note, b does not vary with X_{VdW} because it just depends on the protein-ligand interaction energy.

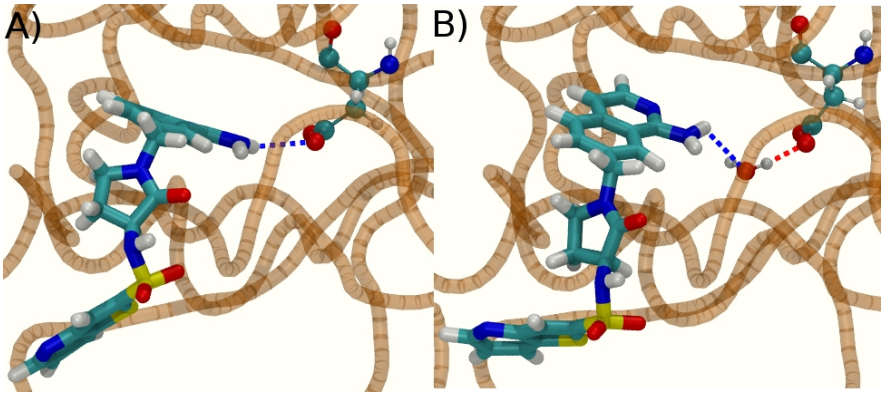


Figure 5.10: Shows how 1F0R ligand changes its binding mode between the A) $\boldsymbol{r}_{\mathrm{c}} = full$ and B) $\boldsymbol{r}_{\mathrm{c}} = bb$ case.

The published Abel *et al.* [132] results were not reproduced while following all simulation procedures and using clustering at the same density cutoff of 2x bulk concentration albeit clustering was done on the grid in the work here while in the Abel work was done from the trajectory. Various alternative clustering parameters were also considered with density clustering at different relative densities to bulk (0.5 to 5 at intervals of 0.5) and a cluster size of 1 Å, as shown in Figure 5.8. Poor correlation is found for all tested clustering parameters. Figure 5.8 shows that the results are insensitive to the density clustering parameters. The ab intio descriptor consistently gave poor R^2 and PI values in contrast to the value obtained in the Abel *et al.* paper [132] quoting an R^2 value of 0.64. Calculations were also repeated without energy minimising the binding poses, with further minimising, and also with protonation states consistent with Abel *et al.* [132], but these efforts did not improve the results (data not shown). These R^2 and PI values are too low to be of use in identification of trends in the FXa system.

5.3.1.3 Evaluation of relative PSAPO energetics

The predictive value of the PSAPO model was then investigated. The PSAPO model (model a) was found to show better R^2 and PI values than the PSAPO-Abel model for X_{vdW} between 0-4 Å as shown in Figure 5.9a. In figure 5.9a the predictive value of the PSAPO model is a function of the size of the monitored grid regions analysed. The best result was found for PSAPO at 2 Å (first solvation shell), with $\boldsymbol{r}_{\mathrm{p}} = bb$ giving a $\mathrm{R}^2 = 0.38 \pm 0.07$ and PI $= 0.65 \pm 0.06$. In contrast the best result for PSAPO model in $\boldsymbol{r}_{\mathrm{p}} = full$ is also at 2 Å giving $\mathrm{R}^2 = 0.12 \pm 0.004$, and PI $= 0.34 \pm 0.004$. This is in contrast to results shown by Abel *et al.* [132] in which the simulation used for the PSAPO-Abel model had all heavy-atom restraints. The improvement based on the restraint protocol $\boldsymbol{r}_{\mathrm{p}} = bb$ suggests that appropriate sampling of the binding site requires side chain rearrangements.

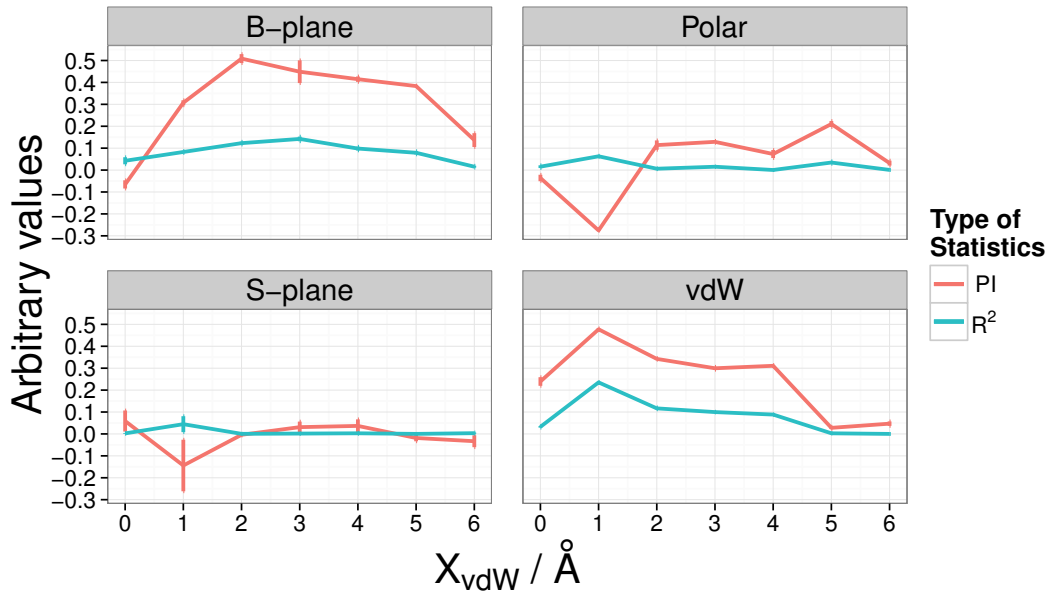


Figure 5.11: The predictive power of the four HOLO variant methodologies, B-plane, Polar, S-plane and vdW (described in Figure 5.5 and section 5.2.5) as a function of the size of the monitored grid region.

5.3.1.4 Evaluation of protein-ligand interaction energetics

Finally, the protein-ligand interaction energy (model b) was investigated (Figure 5.9b). The most reproducible interaction energies come from simulations using the $\boldsymbol{r}_c = full$ conditions which gave the $R^2 = 0.06 \pm 0.05$ and $PI = 0.16 \pm 0.13$. Simulations simulate with $\boldsymbol{r}_c = bb$ conditions gave $R^2 = 0.11 \pm 0.09$ and $PI = 0.20 \pm 0.11$. Both predictive values are poor and unreliable. However, it can be seen that the interaction energies are exaggerated with $\boldsymbol{r}_c = full$ due to no sampling of side chains which would allow water binding. This is seen in the 1F0R ligand case (Figure 5.10) where an aspartate and water tend to interact (5.10B) forming a water bridge which is not present in the *full* restraint case (5.10A) where instead, a direct interaction between the aspartate of a protein and an amine group of 1F0R ligand occurs. Comparison with flexible simulations reveals that the strengths of certain protein-ligand interactions tend to be exagerrated because protein-water coupled motions are removed. In general the magnitudes of the IE-model values decrease in the $\boldsymbol{r}_c = bb$ protocol when compared with those of the $\boldsymbol{r}_c = full$ protocol. It is expected that with longer sampling the $\boldsymbol{r}_c = bb$ could provide a better ranking.

5.3.1.5 Evaluation of relative LIG energetics

The LIG model (model c) also had greatest predictive value at 2 Å distance for the $\boldsymbol{r}_l = full$ case which is shown in Figure 5.9c. This model gave $R^2 = 0.10 \pm 0.004$

and PI = 0.36 ± 0.01 which is comparable to those of the PSAPO-Abel model. This suggests that the FXa relative binding free energies of the present ligand pairs may not be driven by LIG desolvation energetics because of the poor predictive value of the ranking of the ligands.

5.3.1.6 Evaluation of relative HOLO energetics

Again a reminder that the FXa site binders are characterized by an L-shape binding; alignments of all ligands are shown in Figure 5.4. First, HOLO vdW variant 1 described in section 5.2.5 was evaluated and is also shown in Figure 5.9d in the total vdW variant 1 (model d). The results were poorly reproducible primarily because of the water exchanges near the ligands which have an overall large solvent accessible surface area (Figure 5.5 bottom right). The large uncertainties from this solvent exchanging area seem to be linked with the loss of predictive value in the $\boldsymbol{r}_{\mathrm{c}} = bb$ restraint conditions; see Figure 5.11, vdW. However, the poor accuracy in the $\boldsymbol{r}_{\mathrm{c}} = full$ seems to be not a problem of precision but indicates either that the restrained hydration states being sampled are not the equilibrium hydration states being sampled or that the HOLO energetics are not the major reason the different FXa ligands show different binding energies.

After that, instead of focusing on the entire ligand, the polar atoms (HOLO VDW variant 2) grid selection method was investigated (see figure 5.5 top right). This again provided even less predictive value; see Figure 5.11 Polar, suggesting that polar atom differences do not drive HOLO energetic ranking. A final analysis was attempted where a partition of the grid region was done so that a solvent-exposed plane and buried plane were defined from a structural alignment, as shown in Figure 5.5, and Figure 5.4, and as explained in section 5.2.5. In other words, anything "below the plane" toward the protein was not considered. The S-plane method (Figure 5.11) again gives little predictive value. However, the B-plane method, where only waters which are well coordinated between the protein and ligand are considered, gave a slight improvement in the R^2 and PI values (Figure 5.11). This could be because there is better discrimination of hydration sites due to less mobility in a more buried site. Also, the solvent-exposed regions are less likely to be converged and more noisy due to the high mobility and number of water molecules. This shows that when GCT is used the mobility of the water increases the amount of sampling required to get converged results.

5.3.1.7 Multiple descriptor models and conclusions on Factor Xa

After evaluating the multiple descriptor combinations of the 14 models shown in Figure 5.9, the best model was the PSAPO model (at distance cutoff of 2 Å, $\boldsymbol{r}_{\mathrm{p}} = bb$) with

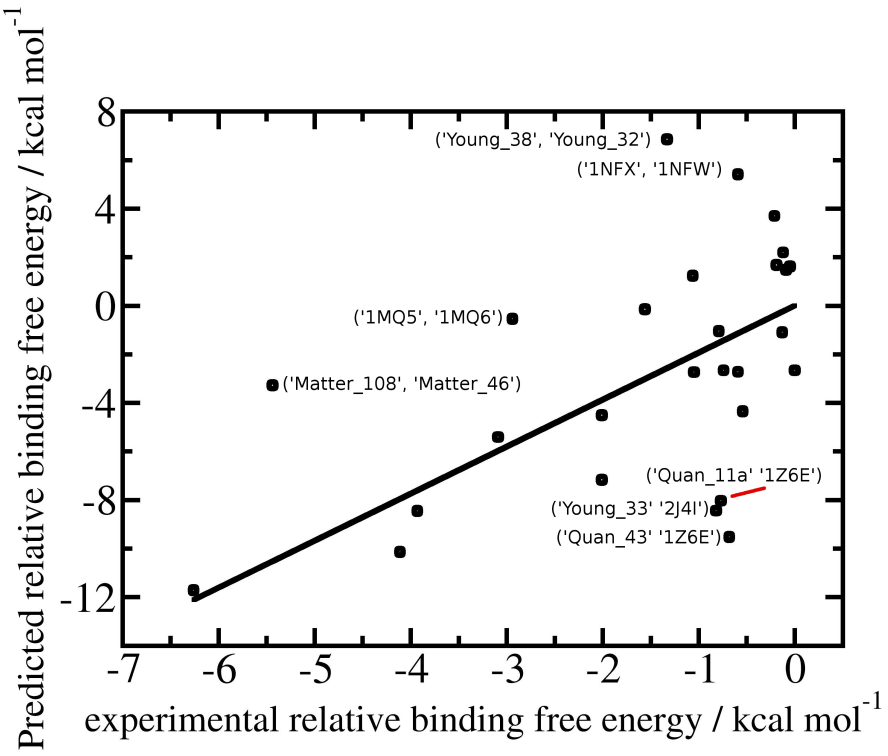


Figure 5.12: Linear regression of the PSAPO ($\boldsymbol{r}$ = *full*, at a distance cutoff of X_{VdW} = 2 Å). The predicted relative binding affinity (y axis) is plotted against the experimental relative binding affinity (x axis) and outliers are also labelled.

an R^2 of 0.38 ± 0.07 and PI of 0.65 ± 0.06. Combinations containing PSAPO have similar predictive power due to the PSAPO term. Overall, the FXa work shows the importance of flexibility; in all models analysed $\boldsymbol{r}$ = *bb* restraints protocols did show greater predictive power.

The results from the best PSAPO model at $\boldsymbol{r}_{\mathrm{p}}$ = *bb* are plotted in figure 5.12 with a few outliers annotated. The pairs of outliers ("Young_38", "Young_32"), ("Young_33", "2J4I") and ("Quan_43", "1Z6E") have one common feature which is the mutation of large functional groups. PSAPO outliers are all caused by large differences in the size of grids analysed. These are caused by large differences in mutated functional groups.

Another avenue which can be pursued is greater sampling of the protein-ligand interaction energies with $\boldsymbol{r}_{\mathrm{c}}$ = *bb* conditions which are still not well converged (errors in Figure 5.9b). This is clear from inspection of the magnitude of the error bars. There is no clear consensus on whether the HOLO energetics with the current amount of sampling are practical. The present results suggest that only well converged grid areas (near the binding site) should be considered due to prohibitive amounts of sampling required to converge water energetics at solvent exposed areas. LIG energetics are better converged and the term is most predictive at a X_{vdW} at 2 Å. Overall, the results favour the use of $\boldsymbol{r}$ = *bb* conditions and indicate a greater predictive power of the APO model

over LIG and HOLO energetics. Nevertheless the top performing model is insufficiently predictive for prospective applications. This may indicate that conformational entropy, strain energy, and other factors may play a role here.

5.3.2 Heat Shock Protein 90

The heat shock protein binding site is quite different from the FXa binding site. It is more buried and less solvent exposed, as shown in Figure 5.13. There are also a few buried waters, typically two conserved waters, and less variability between ligands with the scaffold analysed in this work (Figure 5.13).

5.3.2.1 Convergence of hydration energies

For the HOLO, APO, and LIG relative energetics all ligands are considered relative to ligand **1** and grid points were selected using the vdW methods with all atoms. Inspection of the HOLO values in $\boldsymbol{r}_{\mathrm{c}} = full$ conditions indicates that there is a divergence in energetics after the X_{vdW} cutoff is greater than 5 Å (see Figure 5.14). This seems to be due to a slight shift of the binding pose of ligand **1** with respect to other ligands. This means that as the distance cutoff increases, the other protocol select more bulk grid points at lower cutoff values. Thus with the current protocol only X_{vdW} values from 0 to 5 Å are reasonable. Under the $\boldsymbol{r}_{\mathrm{c}} = bb$ conditions there is a smaller jump in the relative HOLO energetics because the binding pose of **1** and the other ligands are more similar. As well as this waters can sample more regions of the binding site with the $\boldsymbol{r}_{\mathrm{c}} = bb$ protocol.

As for the FXa system, the APO energetics are well converged and have low errors with both restraint conditions (see Figure 5.14). The relative LIG energetics seem to suffer

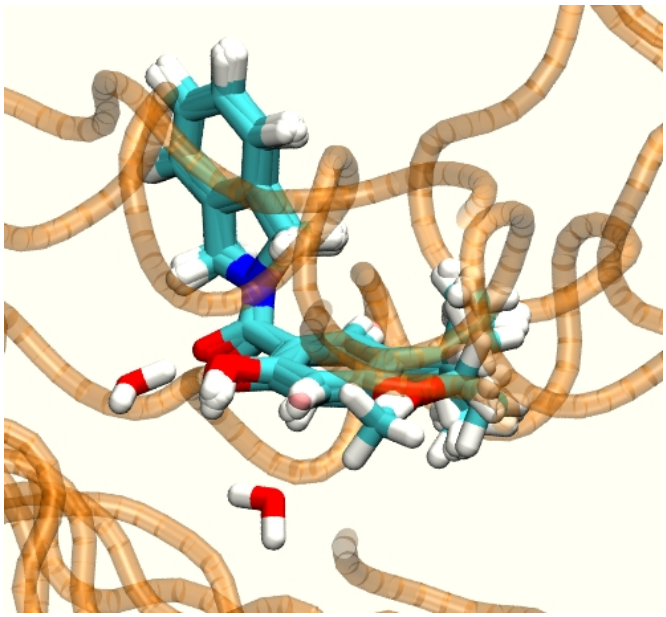


Figure 5.13: Superposition of all the ligands in the HSP90a binding site with two conserved waters. In the tube representation the backbone of the protein is shown.

from drift again, probably due to sensitivity to the enthalpy bulk parameter which suggests a maximum X_{vdW} of 5 Å similar to where the divergence is seen in the HOLO energetics (Figure 5.15). The noisest term is again the HOLO term. This is followed by the LIG term with the PSAPO being the least noisy.

5.3.2.2 Evaluation of relative PSAPO energetics

The APO model (model a) energetics results show strongly anticorrelated ranking in both restraint protocols as shown in Figure 5.16a. There are strong correlations in the R^2 values at larger distances of 5-8 Å X_{vdW} with $\boldsymbol{r}_{\mathrm{p}} = bb$ conditions, but this seems to be more a function of the shape of the vdW surface of the ligands and it is unclear why including regions further away improve correlation. With the $\boldsymbol{r}_{\mathrm{p}} = full$ conditions there is only a strong correlation at $X_{\mathrm{vdW}} = 0$ Å which is a model which has more similarities with the PSAPO-Abel term described in FXa discussion. Overall, the best result is seen at a X_{vdW} of 7 Å which results in $R^2 = 0.83 \pm 0.02$ and PI = -0.92 ± 0.03, but it is unclear why such a large cutoff improves correlations. Below the 5 Å cutoff the best cutoff is at 4 Å for the X_{vdW} which results in a $R^2 = 0.62 \pm 0.01$ and a PI = -0.78 ± 0.02 which still indicate a strong inverse correlation.

The reason for the anticorrelation could be a relationship between the protein desolvation descriptor and other solvation mechanisms. A possible explanation may be a result of the PSAPO model at larger X_{vdW} cutoffs, which can act in two ways. First, when $X_{\mathrm{vdW}} = 0$ Å it is representing a protein desolvation cost but at higher X_{vdW} (when $X_{\mathrm{vdW}} > 0$) Å it acts as a HOLO energetic estimator. However, in the current implementation it is still considered a desolvation cost because the thermodynamic cycle expects the HOLO term to compensate. However, further analysis is required to ascertain exactly why the anticorrelation is present.

5.3.2.3 Evaluation of protein-ligand interaction energetics (IE)

Results for the IE model (model b) of both the $\boldsymbol{r}_{\mathrm{c}} = bb$ protocol and $\boldsymbol{r}_{\mathrm{c}} = full$ are shown in Figure 5.16b with the $full$ protocol having more predictive power. The IE model with $\boldsymbol{r}_{\mathrm{c}} = full$ gives $R^2 = 0.31 \pm 0.04$ and a PI = -0.60 ± 0.06. There is again an anticorrelation which may reflect the dominant influence of binding site waters on the protein and ligand interactions.

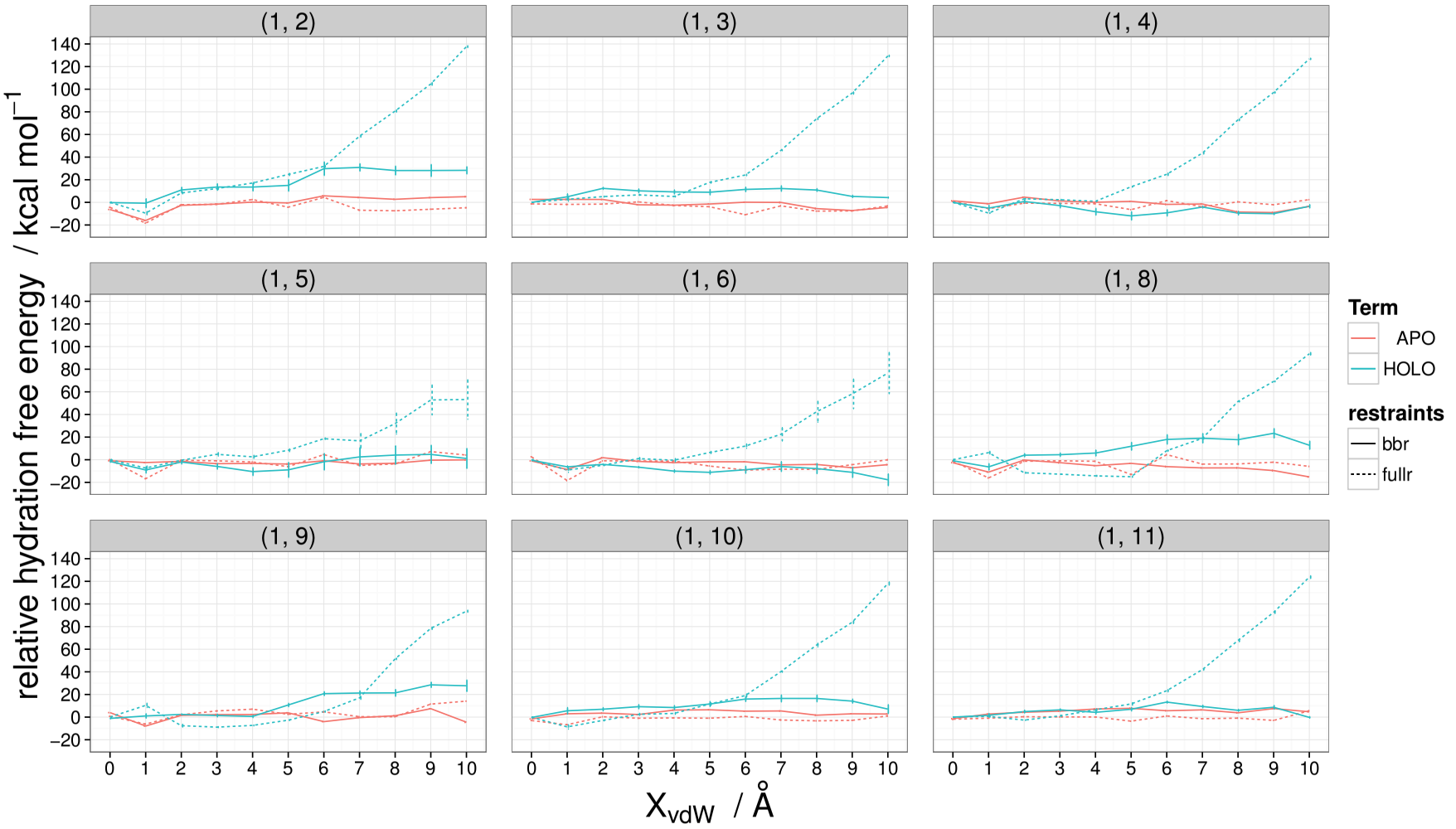


Figure 5.14: The computed APO and HOLO relative hydration free energies as function of the size of the grid region. The HOLO and APO energies are shown as the difference in hydration free energies of ligand 2 minus ligand 1.

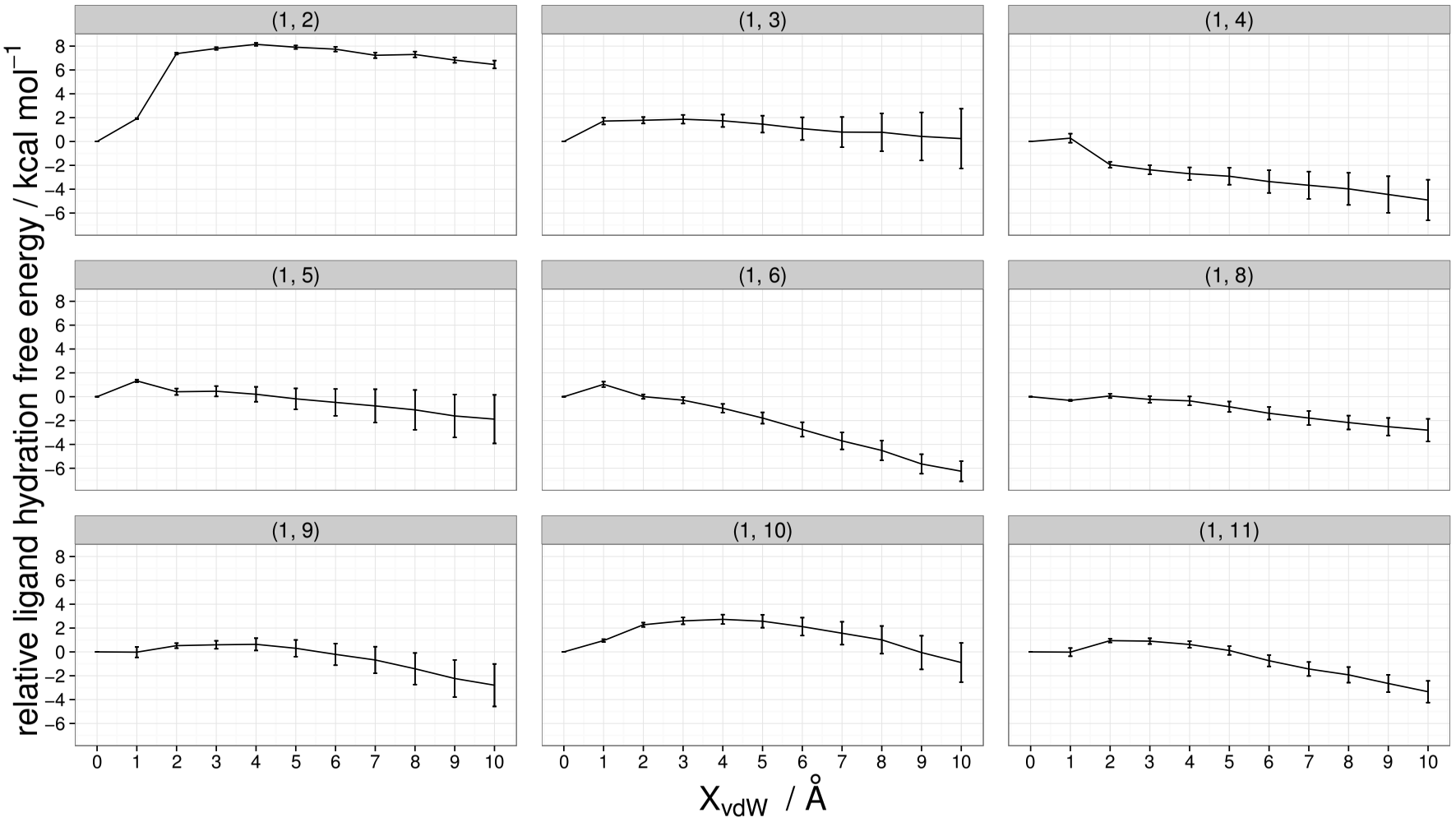


Figure 5.15: The computed LIG relative hydration free energies as function of the size of the grid region. The LIG energies are shown as the difference in hydration free energies of ligand 2 minus ligand 1.

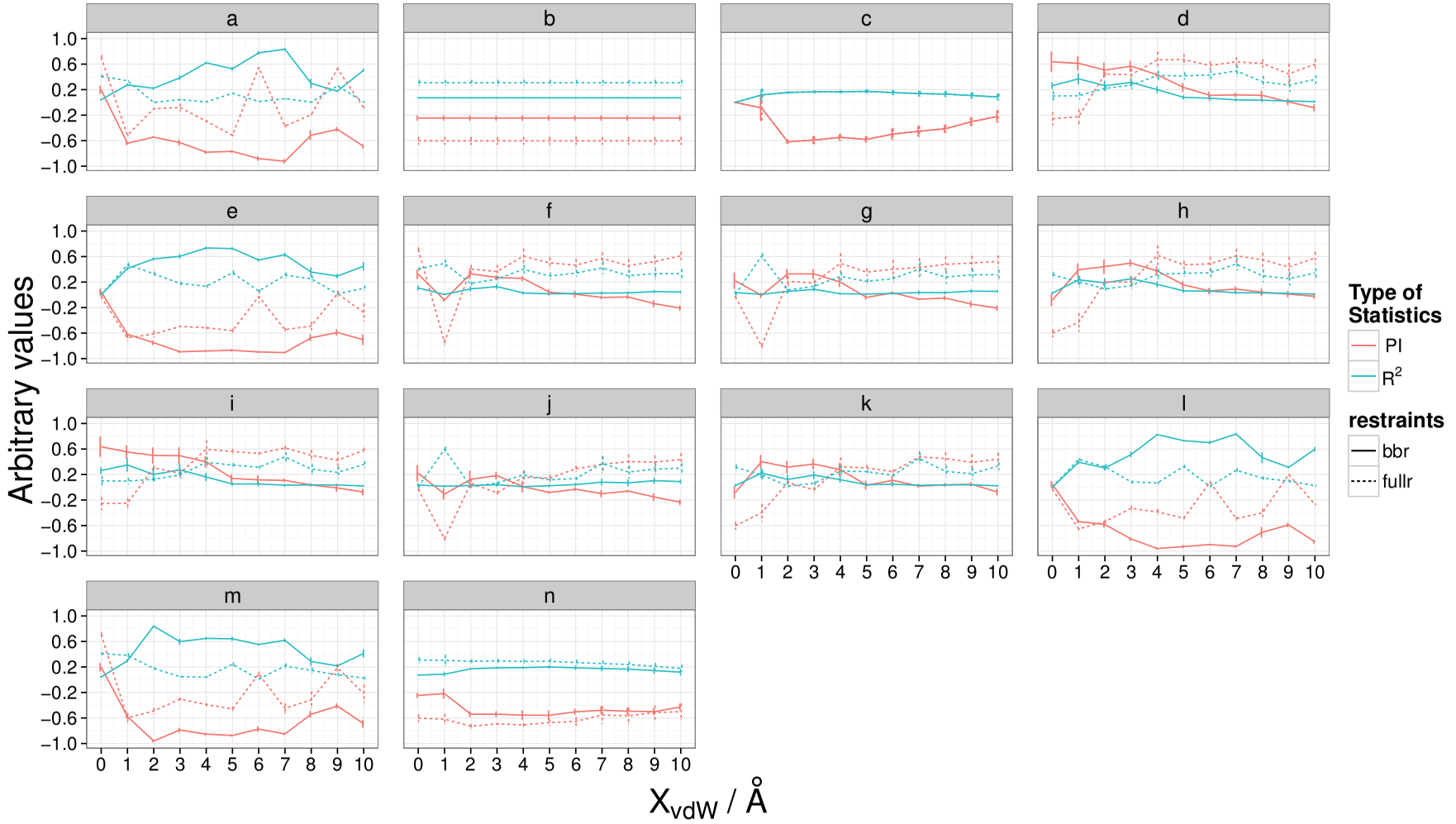


Figure 5.16: The predictive power of all 14 models as a function of the size of the monitored grid region. Labels represent model as shown in table 5.1. Model b does not vary with X_{VdW} because this model depends on the protein-ligand interaction energy. Triplicate simulations (21000 snapshots) are used to compute the standard error of the mean.

5.3.2.4 Evaluation of relative LIG energetics

The LIG model (model c) does not have a good correlation as is shown in Figure 5.14c. The LIG model performs the best at $\boldsymbol{r}_{\mathrm{l}} = full$ when X_{vdW} = 5 Å, giving R^2 = 0.17 ± 0.03 and PI = -0.58 ± 0.04. There is again an anticorrelation which may relate with the incomplete thermodynamic cycle. However, it is clear the anticorrelation is smaller than that of PSAPO, PI = -0.92 ± 0.03, which suggest that protein desolvation is probably a more important driver in comparison to ligand desolvation.

5.3.2.5 Evaluation of relative HOLO energetics

The predictive power of the HOLO model (model d) changes depending on restraint conditions; results are shown in 5.14d. In the $\boldsymbol{r}_{\mathrm{c}} = bb$ protocol, predictive power is greater when the X_{vdW} cutoff is from 0-3 Å. At a higher distance cutoff the $\boldsymbol{r}_{\mathrm{c}} = full$ protocol has better correlations. This may reflect how restraint conditions affect solvent behaviour. The best predictive values are found with $\boldsymbol{r}_{\mathrm{c}} = bb$ when X_{vdW} = 1 Å giving an R^2 = 0.37 ± 0.08 and a PI = 0.61 ± 0.11.

5.3.2.6 Multiple terms models

The GCT data on HSP90a seem to suggest different driving forces for binding compared with the FXa binding site. First of all, hydration thermodynamics seems to play a much larger role because there are buried waters conserved throughout all simulations with both $\boldsymbol{r} = full$ and $\boldsymbol{r} = bb$ in the system. Another major indicator is the large R^2 values obtained from both the APO and LIG terms shown in Figure 5.16. However, the PI statistic presents a problem with ranking which is mostly inversely correlated. This is very strange, unexpected behaviour. This indicates that the protein and ligand desolvation costs seem to be inversely correlated due to overcompensation in the binding site. This could reveal issues with the thermodynamic cycle which assumes complete protein desolvation and ligand desolvation. However, in the real system only partial dewetting of the binding site occurs in most systems. A better protocol may focus on regions clearly desolvated after the binding event. Also, the dynamics of the dewetting process is never thoroughly investigated because the complete binding process is never simulated.

Results show that the best combination is the APO ($\boldsymbol{r}_{\mathrm{p}} = bb$ at X_{vdW} = 2 Å) and LIG ($\boldsymbol{r}_{\mathrm{l}} = full$ at X_{vdW} = 2 Å) model which gives R^2 = 0.841 ±0.004 and PI = -0.961 ±0.013. This is seen in the linear regression (Figure 5.17), where some outliers are shown. The largest discrepancy is in the relative free energy estimate between **1** and **9** which was found to have a predicted binding affinity which was too negative.

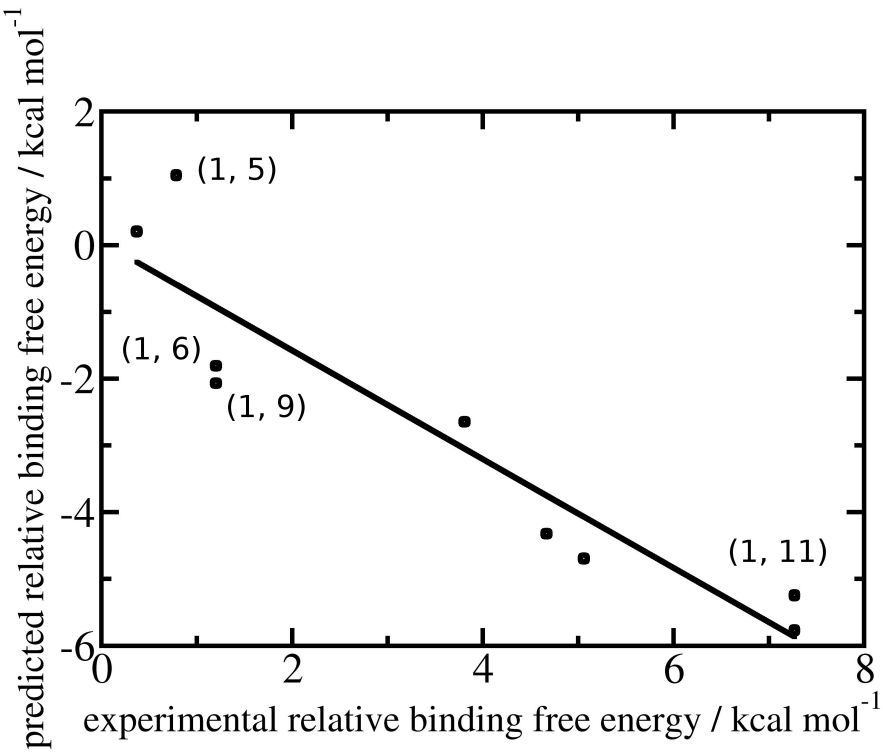


Figure 5.17: Linear regression of the best combination *LIG* ($\boldsymbol{r}_{\mathrm{l}} = full$ at $X_{\mathrm{VdW}} = 2$ Å) and *APO* (($\boldsymbol{r}_{\mathrm{p}} = bb$ at $X_{\mathrm{VdW}} = 2$ Å). The predicted score is plotted against the experimental relative binding affinity. Outliers are also labelled.

Otherwise the regression inversely correlates quite well with experimental data. The nature of inverse correlation is puzzling and requires further investigation.

5.4 Conclusion

A correct balance between solvent and protein-ligand interaction energies is the key to predicting a molecular recognition event of similar congeneric ligands to a particular protein system of interest. Obtaining predictive rankings from solvent descriptors extracted from GCT analyses requires a careful selection of restraint protocol. Interaction energies also heavily rely on restraint conditions because weaker restraints may allow conformational changes to optimise interactions with a water (FXa case) which in turn lowers the protein-ligand interaction energy component. Depending on how solvent exposed an area is, it appears that sampling for 10-100 ns, is insufficient to obtain converged data. For this reason it seems pragmatic to focus grids on regions where solvent exchange is slow. To further improve analyses a thorough sampling of a simple system in a completely unrestrained simulation would provide a good model. Each conformation could then be clustered with an RMSD or alternative method so that hydration thermodynamics at each particular conformational cluster can be later

Boltzmann weighted by the probability of the conformational state. In this way the effect of flexibility could be more thoroughly assessed in GCT. As well as these improvements, inclusion of conformational entropy and strain energy (internal energy cost upon binding) could improve binding affinity predictions.

Several aspects of the binding event have been included in GCT analyses. These include the solvation of polar and charged groups. Hydrophobic interactions are estimated by MD interaction energies from the force field. However, an important missing aspect is the dynamics which is neglected in this work.

Chapter 6

Hydration thermodynamics of a diverse dataset of druggable proteins

6.1 Introduction

Water plays a crucial role in the structure and dynamics of proteins. It is implicated as a mediator of forces between different protein surfaces [139], and also thought to be very important for understanding the protein folding process, where the main driving force is thought to be the burial of hydrophobic side chains of amino acids [140]. Understanding how water interacts with protein structure and relates to protein function is important for enzyme catalysis, molecular recognition of various events including protein-DNA [141], protein-protein [142] and protein-ligand interactions [143]. If these forms of interactions can be understood at the molecular level, new avenues for the creation of novel therapeutics can be opened.

The work reported here was mostly inspired by the work of Beuming *et al.* [144] where a large dataset of 27 proteins was tested using the Watermap software created by Schrödinger to investigate general water thermodynamics around structural motifs found in proteins using IFST. It was also used to predict the location of binding sites on proteins. The dataset included many famous drug targets such as: HMGÂŋCOA reductase (statins for cholesterol reduction), PDE5 (Viagra), and cyclooxygenase (Aspirin). As well as those the cancer targets: caspase1, MDM2, CDK, cAbl tyrosine kinase; the blood disease target: thrombin; two HIV reverse transcriptase structures, HIV integrase; Flu target: neuraminidase; and the antibiotic target PBP (penicillin binding protein) are all included. The study here covers a slew of the aforementioned

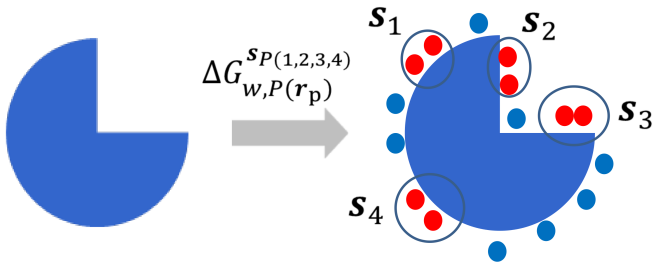


Figure 6.1: Evaluation of hydration energies of a region **s** which can be residues or pockets of a protein, *P*. Proteins are depicted by blue shapes. In all GCT analyses, water molecules (red circles) are inside the monitored regions, $\mathbf{s}_{P(1,\ldots n)}$, contribute to the computed hydration free energies, whereas those that are out of the monitored regions in blue are ignored. Only restraint protocol ($\boldsymbol{r}_\mathrm{p} = full$) is used to control the protein (p).

proteins to further compare the IFST (Watermap) results to the GCT implementation called *Nautilus*, investigate novel analyses on how electrostatics calculations compare with GCT, and also to analyse the average pocket hydration properties to that of a binding site hydration properties. Electrostatics calculations are compared to develop an understanding if there is any correlation between the electrostatics and water hydration thermodynamics. The analysis where hydration properties of average pockets are compared to an average binding site is used to see if a typical binding site differs in terms of its hydration properties. Knowledge of these properties could help drug designers have better grasp of typical hydration environments found in binding sites.

6.2 Theory

In this work the thermodynamic properties of water around amino acids, density-clustered hydration sites, and pockets were investigated in the dataset of druggable proteins. Various other analyses were also implemented, including the comparison of density-clustered hydration sites with a Poisson-Boltzmann electrostatic calculation.

6.2.1 GCT, localised protein hydration energy

The equivalent of the protein hydration energy calculation was run for different areas of each protein. For example regions 1,2,3, and 4 in Figure 6.1 would have a free energy of hydration, for each respectively, $\Delta G^{\mathbf{s}_{P(1)}}_{w,P(\boldsymbol{r}_\mathrm{p})}$, $\Delta G^{\mathbf{s}_{P(2)}}_{w,P(\boldsymbol{r}_\mathrm{p})}$, $\Delta G^{\mathbf{s}_{P(3)}}_{w,P(\boldsymbol{r}_\mathrm{p})}$, $\Delta G^{\mathbf{s}_{P(4)}}_{w,P(\boldsymbol{r}_\mathrm{p})}$. Regions typically included pockets or density-clustered hydration site or an area around a particular residue of interest, shown in Figure 6.1. Grid cell theory calculations are then run for particular regions of interest, $\mathbf{s}_{P(1)}$, $\mathbf{s}_{P(2)}$, and $\mathbf{s}_{P(\ldots n)}$. The GCT method calculates hydration free energy using the *Nautilus* protocol which is identical to the methods of the previous chapters.

6.2.2 APBS, Adaptive Poisson-Boltzmann Solver

It was of interest to see if the magnitude of the electrostatic potential at a particular region of space would correlate with the thermodynamic stability of the GCT hydration free energy computed. To enable a reasonable comparison, only density-clustered sites obtained from simulations were assessed with Poisson-Boltzmann calculations. This effectively discards regions of space that have a high electrostatic potential, ϕ, but are not solvent accessible. Note that this should not be confused with the electrostatic field, $\vec{E}$, whose vector components are the negative derivatives of the electrostatic potential in terms of each vector component, shown as follows: $\vec{E} = -\nabla\phi$.

The APBS method was developed by Baker *et al* [145]. It is based on the Poisson-Boltzmann equation (PBE):

$$-\nabla\cdot\epsilon(r)\nabla\Phi(r) + \kappa^{-2}(r)\text{sinh}\Phi(r) = f(r) \tag{6.1}$$

which is a second-order nonlinear elliptic partial differential equation (nonlinear PBE), relating the *dimensionless* electrostatic potential, Φ, to the dielectric properties of the solute and solvent, ϵ, to a measure of the ionic strength and accessibility of these ion into the solute, κ^{-2} and the distribution of the solute atomic partial charges, f. The *dimensionless* electrostatic potential, Φ, relates to the real electrostatic potential, ϕ (J/C) as follows [146]:

$$\Phi(r) = q\phi(r)/(kT). \tag{6.2}$$

In eqs 6.1 and 6.2, r is the position of the point charge. In eq 6.2 q, is the elementary charge (C), and kT gives energy (J). This equation can be linearised into the linearised PBE (LPBE) by approximating that $\text{sinh}\phi(x) \approx \phi(x)$. The next advance in the APBS method is the domain discretisation of the problem so that the equation can be solved in parallel over many processors, which enables application to large systems.

This solver uses a parallel focusing algorithm which is composed of the following steps:

1. First, a coarse resolution solution of the entire problem is computed by each processor involved.

2. This approximate solution is used to partition the problem into P subdomains assigned to P processors

3. Each processor then solves a fine-scale finite difference calculation on the domain and a small overlap region outside.

4. After the fine-scale calculations are complete a master processor accumulates the desired data from the other processors and assembles individual results into the global solution which has been proven to be as rigorous as the solution on a single

big mesh.

6.3 Simulation protocols

6.3.1 Preparation of proteins

All of the following 17 PDB [147] structures were kept in the dataset (1BMQ, 1E1X, 1E66, 1E9X, 1EZQ, 1HWL, 1HWR, 1IEP, 1KV1, 1M17, 1NLJ, 1OYN, 1PTY, 1QMF, 1UDT, 4COX and 1YCR). Essentially the selected proteins were prepared in the same way as the PSAPO structure in section 5.1.4. The structures were used for PSAPO (protein alone) simulations after the respective ligands was removed (if there were dimers or homodimers only the relevant monomer was used). After the ligand was removed, tleap (AMBER 11 [84]) was used to generate protein parameters from the AMBER99SB forcefield [135]. The protein was solvated with TIP4P-EW [81] waters in a rectangular box. The edges of the box extended at least 11 Å away from the edges of the protein. Respective disulfide bonds for each protein followed bonding shown in their respective PDB files. The system was first minimised and equilibrated for 1 ns in AMBER before the production run.

6.3.2 Production run

All molecular simulations were produced using the software Sire/OpenMM by linking the general purpose molecular simulation package Sire (revision 1786), with the GPU molecular dynamics library OpenMM (revision 3537) [89]. Simulations were run at a pressure of 1 atm and temperature of 298 K using an atom-based generalized reaction field nonbonded cutoff of 10 Å for the electrostatic interactions [51], and an atom-based nonbonded cutoff of 10 Å for the Lennard-Jones interactions. A velocity-Verlet integrator with a time step of 2 fs was used. Temperature control was achieved with an Andersen thermostat with a coupling constant of 10 ps^{-1} [48]. Pressure control used attempted isotropic box edge scaling Monte Carlo moves every 25 time steps. The OpenMM default error tolerance settings were used to constrain the intramolecular degrees of freedom of water molecules. For each protein system one simulation of 50 ns were run with a random velocity assignment. Snapshots were stored every 1 ps and were written in a DCD format. The first ns of each trajectory was discarded to insure the system was well equilibrated prior to sampling for all simulations. All proteins were restrained using $\boldsymbol{r}_{\mathrm{p}} = full$ heavy-atom, positional, harmonic restraints with a force constant of 10 kcal $mol^{-1}Å^{-2}$.

6.3.3 Grid placement

Rectangular grids were placed so that they extended to at least 3.8 Å away from the extreme edges of each protein. This cutoff was deemed sufficient to capture short ranged interactions.

6.4 Analyses

6.4.1 Amino acid analyses

A simple distance cutoff of 4 Å was used to allocate regions near particular amino acids throughout the dataset, identical to the cutoff used by Beuming *et al* [144]. However, in the Beuming *et al.* [144] study, this cutoff is only used to pick out density-clustered sites, whereas here this is not done because a general analysis using all grid points was preferred. This is because the hydration free energies of low density water sites were also included which could also be vital in understanding hydration behaviour of proteins. Each area around the amino acid was treated as a separate site and data were then collected for each amino acid. A distribution for each amino acid was generated per site which was then normalised to yield per-water statistics. In all cases, not all atoms of the amino acids were investigated but only regions around functional groups of their side chains. These group were the carboxyl groups of the aspartates; the nitrogens of the lysines and arginines; the hydroxyl groups of threonine, serine and tyrosine; amides of glutamine and asparagine; the ring atoms of tyrosine, phenylalanine and tryptophan; and the aliphatic atoms of leucine, isoleucine, valine, and alanine. However, as well as this, the hydroxyl of the tyrosine, as well as the sulfur of the methionine, are also treated as separate groups.

These were then compared using a nonparametric statistic to compare distributions, called the Kolmogorov-Smirnov test, which measures the likelihood that two distributions were derived from the same distribution.

The Kolmogorov-Smirnov test measures the distance between the empirical cumulative distribution function of a sample compared to the cumulative distribution function of a reference distribution. Because it is nonparametric and sensitive to the location and shape of the empirical cumulative distribution, it is a robust way to compare two samples. The empirical cumulative distribution function ($F_n(x)$) is given by equation 6.3:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[-\infty,x]}(X_i) \tag{6.3}$$

where n observations are binned by the indicator function $I_{[-\infty,x]}$ which is equal to 1 if $X_i \leq x$ otherwise it is equal to zero. This procedure is repeated for both datasets and then a Kolmogorov-Smirnov statistic (D_n) is computed as follows:

$$D_n = sup_x|F_n(x) - F(x)| \tag{6.4}$$

where sup_x is the supremum, or lowest upper bound of the set of distances derived from the two empirical cumulative distribution functions. This statistic can range from zero to one.

6.4.2 Density clustered sites

Density clustered sites were calculated in the same way as in previous chapters, with a neighbour cutoff set at 1.5 Å and a density threshold of at least 1.5× that of bulk water. Also, sites 10× more dense than bulk were analysed separately with the same neighbour cutoff of 1.5 Å.

6.4.3 Crystallographic water analysis

A simple test was done to compare density-clustered sites with crystal-water sites from experiment. This test consisted of the following steps:

1. The crystal waters, and protein from the PDB, is aligned to the simulation frame of reference.

2. A density clustered grid was produced to obtain clustered sites from simulation data.

3. For each crystal water, the minimum distance to a density clustered site is calculated.

4. The minimum distance of the crystal water site and the density of the site are used to generate the analysis.

6.4.4 Comparing pockets and binding sites

In this analysis a simple comparison of the average hydration thermodynamic properties of the top 10 druggable pockets as found by fpocket [148], are compared to those of the actual binding site. Fpocket works by using the concept of alpha spheres [149]. Alpha spheres are defined as spheres which must contact at least 4 atoms with an

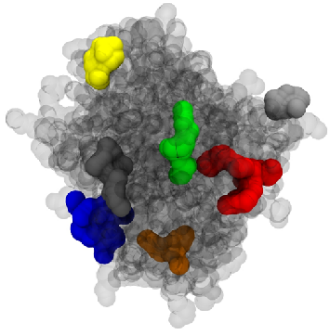


Figure 6.2: Coloured pockets found in the PDB structure 1E1X using the fpocket software [148].

identical distance from the alpha sphere centre. These alpha spheres in turn reflect the local curvature defined by the atoms. In a protein, binding sites tend to be occupied by larger quantities of small radii alpha spheres within the protein, while the exterior is typically composed of larger radii alpha spheres, finally intermediate radii seem to reflect more exposed binding sites and clefts. The radii do vary between hydrophobic and hydrophilic regions of the protein. Thus, one can investigate alpha spheres that are generated, and clusters of alpha spheres can be located. This search is done thoroughly by treating alpha-sphere centres as Voronoi vertices. Fpocket is implemented with the following methodology in brief:

1. Voronoi tesselation and alpha sphere detection: Distance between Voronoi vertices (alpha sphere centres) are found and a maximum and minimum size of alpha spheres are chosen. Only tightly packed alpha spheres are chosen which reflect tight atom packing. Alpha spheres are then labelled by whether they contact 3 apolar atom or 2 or more polar atoms where they are then defined as apolar or polar respectively.

2. These alpha spheres are then clustered further with first a distance criterion based on the neighbour lists. The centres of mass of the clusters can then be used to define larger clusters which can be merged with others if they are close in proximity. Filtering based on minimum numbers of apolar and polar alpha spheres can then be done to fine-tune the analysis.

3. These are then characterised using pocket descriptors which are: the number of alpha spheres; mean local hydrophobic density; apolar proportion of alpha spheres; polarity score (polarity overall all amino acids involved in the pocket, 1 for polar 0 for apolar); and finally the alpha sphere density.

An example of an output of fpocket is shown in Figure 6.2 where pockets for the 1E1X protein structure are visualised. The general workflow is as follows:

1. Use fpocket to generate the top 10 druggable pockets for each protein in the dataset.

2. Fpocket produces a pdb or pqr file which contains coordinates representing the pocket.

3. Using these coordinates all grid points within 1 Å of any pocket-site coordinates define an allocated region.

4. The free energy of hydration of the resulting region is computed.

With this pocket-analysis protocol, each pocket has defined parameters that include ΔG_{X+w}, ΔH_{X+w}, $-T\Delta S^{\circ}_{X+w}$, relative density, average number of waters, and the volume of the site.

The properties of known ligand-binding sites were extracted. Fifteen complete binding sites were found in the dataset with one binding site found in each 4COX, 1E1X, 1E66, 1E9X, 1EZQ, 1IEP, 1KV1, 1M17, 1NLJ, 1OYN, 1UDT and two binding sites in 1PTY and 1QMF. Only complete smallÂŋmolecule binding sites were considered (1YCR was excluded since it had a peptide binder). Grid points near the coordinates of the ligand in its bound conformation are allocated if they are within 1 Å of the ligand. The statistics between the average pocket, and the binding site are then compared to identify any large differences.

6.4.5 Poisson-Boltzmann electrostatics comparison with the hydration enthalpy of a site

The following protocol was used to implement this analysis:

1. Generate a large coarse grid with APBS but specify that the fine grid contains the same spacing and density of the GCT computed grid (in this case 1 Å grid density and making sure the two grids are aligned).

2. Density sites are obtained from the GCT clustering method (as discussed for crystal water analysis).

3. The average absolute magnitude of the electrostatic potential of the region is averaged and then compared with the enthalpy of the identical region.

In this way one can see if electrostatic potential strength in a particular region correlates with water stability.

Amino acid	$\Delta G^{\mathbf{s}_{P(1\ldots n)}}_{w,P(\boldsymbol{r}_\mathrm{p})}$ (kcal mol^{-1} water^{-1})	$\Delta H^{\mathbf{s}_{P(1\ldots n)}}_{w,P(\boldsymbol{r}_\mathrm{p})}$ (kcal mol^{-1} water^{-1})	$-T\Delta S^{\mathbf{s}_{P(1\ldots n)}}_{w,P(\boldsymbol{r}_\mathrm{p})}$ (kcal mol^{-1} water^{-1})
ALA	-3.79 ± 0.15	-4.10 ± 0.16	0.31 ± 0.02
ILE	-3.36 ± 0.27	-3.72 ± 0.24	0.36 ± 0.05
LEU	-3.55 ± 0.22	-3.67 ± 0.30	0.69 ± 0.21
MET	-2.94 ± 0.17	-3.23 ± 0.18	0.28 ± 0.02
VAL	-3.78 ± 0.20	-4.09 ± 0.21	0.31 ± 0.02
PHE	-4.17 ± 0.19	-4.49 ± 0.20	0.33 ± 0.02
TRP	-4.07 ± 0.28	-4.43 ± 0.30	0.36 ± 0.03
TYR	-4.10 ± 0.14	-4.41 ± 0.15	0.44 ± 0.10
ASP	-6.99 ± 0.19	-7.58 ± 0.20	0.59 ± 0.01
GLU	-6.94 ± 0.14	-7.56 ± 0.14	1.14 ± 0.37
ARG	-5.04 ± 0.12	-5.35 ± 0.13	0.31 ± 0.01
HIS	-5.55 ± 0.54	-6.08 ± 0.56	0.54 ± 0.03
LYS	-5.39 ± 0.14	-5.70 ± 0.13	0.44 ± 0.13
ASN	-4.07 ± 0.16	-4.32 ± 0.17	0.25 ± 0.01
GLN	-3.61 ± 0.15	-3.82 ± 0.16	0.61 ± 0.40
SER	-4.07 ± 0.14	-4.34 ± 0.15	0.26 ± 0.02
THR	-4.81 ± 0.27	-5.08 ± 0.28	0.65 ± 0.39
TYR-OH	-4.04 ± 0.22	-4.34 ± 0.25	0.44 ± 0.14
CYS	-3.25 ± 0.35	-3.62 ± 0.36	0.37 ± 0.05
GLY	-4.25 ± 0.18	-4.55 ± 0.20	0.60 ± 0.31
PRO	-3.41 ± 0.13	-3.76 ± 0.14	0.58 ± 0.24
MET-S	-2.32 ± 0.27	-2.58 ± 0.29	0.25 ± 0.04

Table 6.1: The average enthalpy, entropy and free energy of hydration per water around all the amino acids separated by amino acid type, aliphatic (ALA, ILE, LEU, MET, VAL), aromatic (TRP, TYR, PHE), negatively charged (ASP, GLU), positively charged (HIS, LYS, ARG) and polar (TYR-OH, THR, SER, ASN, GLN) and the rest (CYS, GLY, PRO) is shown. As well as the tyrosine side chain as a whole the hydroxyl group of the tyrosine is looked at separately as well as the sulphur of the methionine.

6.5 Discussion

6.5.1 Amino acid analysis

Different trends can be found depending on the type of amino acids that are in proximity to a water molecule. Here different free energy distributions are seen about

Amino acid groups	$\Delta G^{\mathbf{s}_{P(1\ldots n)}}_{w,P(\boldsymbol{r}_\mathrm{p})}$ (kcal mol^{-1} water^{-1})	$\Delta H^{\mathbf{s}_{P(1\ldots n)}}_{w,P(\boldsymbol{r}_\mathrm{p})}$ (kcal mol^{-1} water^{-1})	$-T\Delta S^{\mathbf{s}_{P(1\ldots n)}}_{w,P(\boldsymbol{r}_\mathrm{p})}$ (kcal mol^{-1} water^{-1})	Ref [144], $\Delta G^{\mathbf{s}_{P(1\ldots n)}}_{w,P(\boldsymbol{r}_\mathrm{p})}$ (kcal mol^{-1} site^{-1})	Ref [144], $\Delta H^{\mathbf{s}_{P(1\ldots n)}}_{w,P(\boldsymbol{r}_\mathrm{p})}$ (kcal mol^{-1} site^{-1})	Ref [144], -T$\Delta S^{\mathbf{s}_{P(1\ldots n)}}_{w,P(\boldsymbol{r}_\mathrm{p})}$ (kcal mol^{-1} site^{-1})
aliphatic	-3.64 ±-0.11	-3.87 ± 0.13	0.45 ± 0.07	1.85 ± 3.20	-0.23 ± 2.96	2.08 ± 1.26
aromatic	-4.12 ± 0.11	-4.44 ± 0.11	0.40 ± 0.06	1.80 ± 2.51	-0.08 ± 2.43	1.88 ± 1.06
carboxylic	-6.96 ± 0.11	-7.57 ± 0.12	0.89 ± 0.21	-1.48 ± 1.81	-3.45 ± 1.84	1.97 ± 1.03
ARG	-5.04 ± 0.12	-5.35 ± 0.13	0.31 ± 0.01	0.45 ± 1.71	-1.49 ± 1.98	1.94 ± 0.97
LYS	-5.39 ± 0.14	-5.70 ± 0.13	0.44 ± 0.13	-0.03 ± 1.51	-1.67 ± 1.75	1.63 ± 0.91
amide	-3.85 ± 0.11	-4.08 ± 0.12	0.42 ± 0.19	1.18 ± 1.82	-0.53 ± 1.82	1.71 ± 0.98
hydroxyl	-4.33 ± 0.12	-4.60 ± 0.13	0.44 ± 0.14	1.05 ± 1.92	-0.80 ± 2.04	1.86 ± 1.04

Table 6.2: The average enthalpy, entropy and free energy of hydration per water around all the amino acids separated by amino acid type, aliphatic (ALA, ILE, LEU, VAL), aromatic (TRP, TYR, PHE), negatively charged (ASP, GLU), positively charged (HIS, LYS, ARG) and polar (TYR-OH, THR, SER, ASN, GLN) and the rest (CYS, GLY, PRO) is shown and compared to results from the watermap study [144]. Note: NA means not available

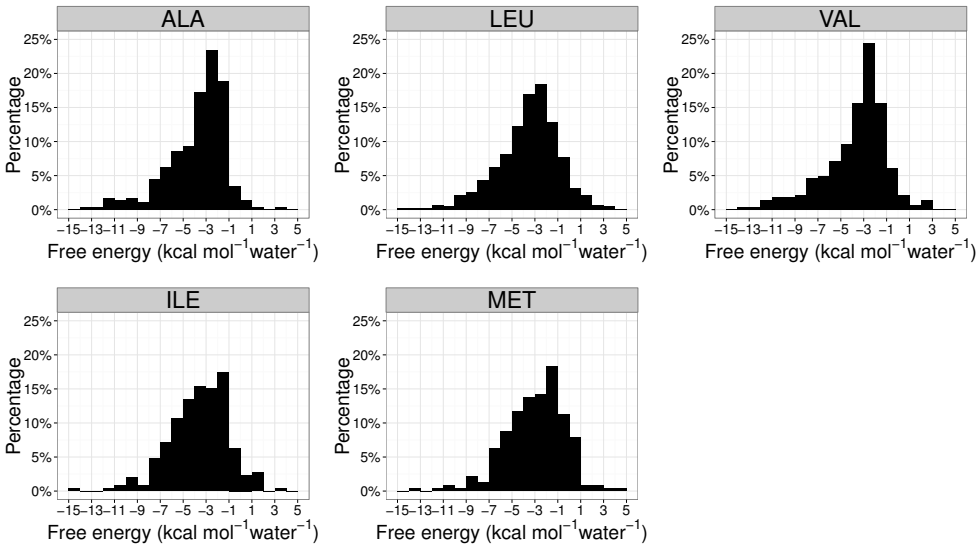


Figure 6.3: Aliphatic amino acids: alanine, isoleucine, leucine, methionine and valine are compared. $\Delta G^{\mathbf{s}_{P(1\ldots n)}}_{w,P(\boldsymbol{r}_{\mathrm{p}})}$ probability distributions of regions surrounding 4 Å of the residue are shown.

various amino acids. The differences between distributions are then assessed using a Kolmogorov-Smirnov (KS) test within amino acid groups: polar, negatively charged, positively charged, aliphatic, and aromatic types of amino acids. As well as this, the average per-water free energy, enthalpy, and entropy of hydration are calculated for each amino acid as shown in table 6.1. When one compares the Watermap values obtained from the Beuming *et al.* work [144] (table 6.2), smaller magnitudes for the enthalpy of hydration are observed in comparison to GCT. This could be because density clustering was not used in the method here to select GCT regions. Enthalpies appear to be much more negative in the GCT case, while entropies are smaller than the IFST entropies calculated by Beuming *et al.* [144]. This difference is caused by the different formulation of the entropies. GCT adopts a molecular view point in cells as opposed to the IFST system view point which relies on molecular distribution functions as discussed by Henchman *et al* [61], and was shown to exaggerate entropies in his study. Another interesting observation is that in GCT, negatively charged amino acids tend to decrease the entropy significantly more than any other amino acid group while in the IFST result there is no large difference in how amino acids decrease the entropy. However, the overall ranking of the amino acids with both methods seems to follow similar trends.

First, aliphatic amino acids were investigated. There is little difference between the free-energy distributions shown in Figure 6.3 except for the methionine whose free-energy distribution deviates the most in statistical tests, as shown in Figure 6.4 (first section of rows). The amino acids which have the most similar free energy distributions are the most similar in size, including alanine, isoleucine, and leucine. The most different

in the group is methionine possibly due to the effect of the sulfur atom in its sidechain.

Aromatic amino acids (PHE, TRP, and TYR) show few differences amongst themselves, which is seen in the average free energy per-water shown in table 6.1. This is also reflected in the comparison of the distributions shown in Figure 6.4 and Figure 6.5.

With the negative amino acids there is a clear stabilisation of waters with average hydration free energies of -6.99, and -6.94 kcal mol^{-1} water^{-1} for aspartate and glutamate, respectively. They both have similar distributions (Figure 6.5) which is reflected in the KS tests as well (Figure 6.4).

Polar amino acids were also analysed. Threonine is the amino acid that stabilises water the most, followed by serine and asparagine as shown in table 6.1 and in Figure 6.5. The side-chain amide-containing amino acids stabilise waters less than hydroxyl-containing functional groups which are also significantly different from each other in the KS tests shown in Figure 6.4.

Positively charged amino acids all have similar profiles, as shown in table 6.1. Arginine and lysine have more similar free-energy distributions, while histidine seems to have a broader distribution (shown in Figure 6.5).

Overall, the negatively charged amino acids seem to decrease hydration free energy the most, followed by positively charged amino acids. All other types of amino acids do not reveal a much greater stabilisation. This implies that the local environment determines the hydration thermodynamics with only charged residues causing a correlated decrease

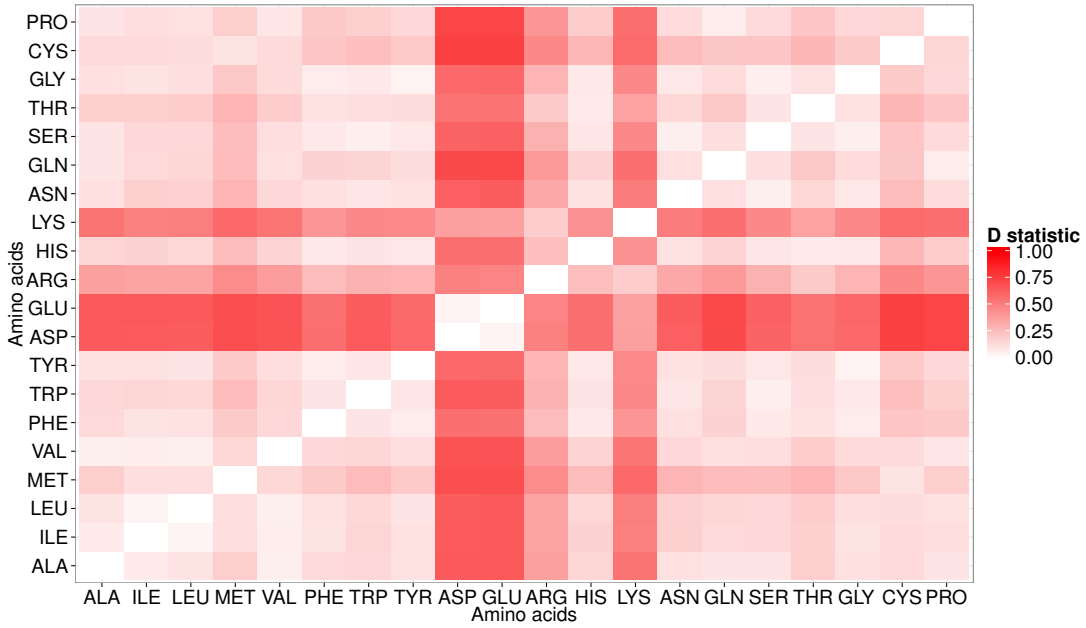


Figure 6.4: Figure provides the Kolmogorov-Smirnov statistics between empirical cumulative distribution functions of amino acid hydration environment (per water hydration free energies).

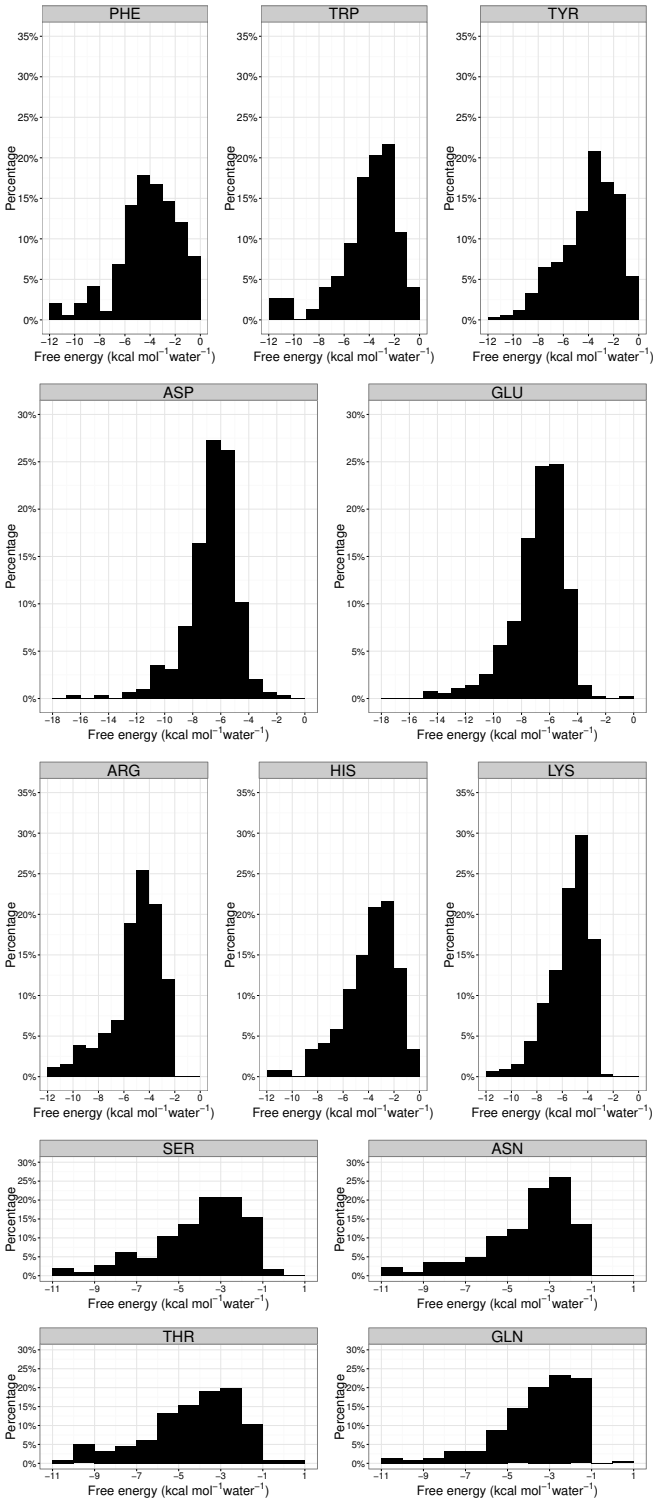


Figure 6.5: Aromatic amino acids: tyrosine, tryptophan, and phenylalanine; negatively charged amino acids: aspartate and glutamate; positively charged amino acids: lysine, histidine and arginine; and polar amino acids: asparagine, glutamine, serine and threonine are compared. $\Delta G^{\mathbf{s}_{P(1\ldots n)}}_{w,P(\mathbf{r}_{\mathrm{p}})}$ probability distributions of regions surrounding 4 Å of each residue are shown.

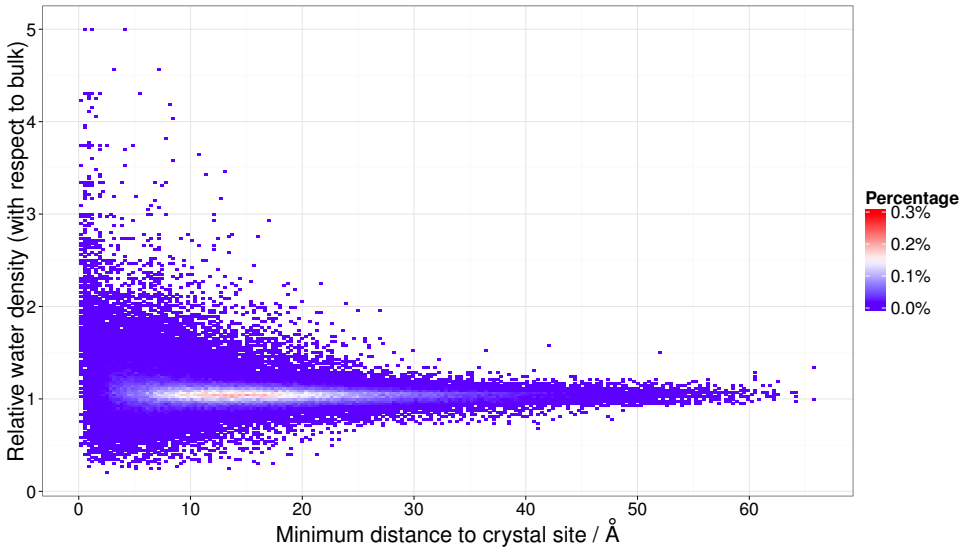


Figure 6.6: Here crystallographic water sites are compared to nearby density clustered water sites. Higher probabilities are shown from red to blue.

in free energy.

6.6 Crystallographic water analysis

Next 1716 crystallographic water sites (derived from all proteins of the dataset except those which did not have any defined including: 4COX, 1BMQ, 1HWR, 1NLJ and 1YCR) were compared to the clusters derived from grid densities computed from molecular dynamics snapshots. Figure 6.6 shows how the density of clustered sites varies as a function of distance to the crystal water site. The figure shows that the further from a crystal site one is, the more likely the cluster site has bulk-like water densities. By

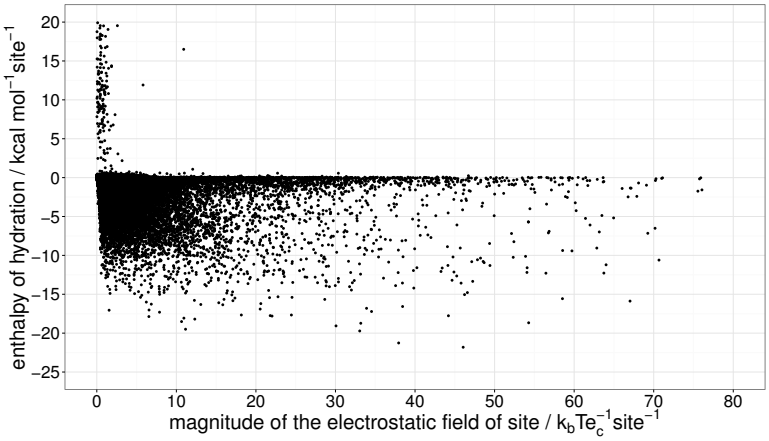


Figure 6.7: The average magnitude of the electrostatic potential of a site is compared with the enthalpy of hydration of a particular density site found.

contrast the cluster sites closer to the crystal-water sites tend to have densities greater than bulk. This is an expected result and shows a slight skew toward higher relative densities at values closer to a crystal site. There are waters which have lower density than bulk as well in the range below 10 Å primarily. These waters are typically less accessible regions on the protein surface. This indicates overall that crystallographic techniques are better suited to discern more dense water sites, rather than low-density water sites which could nevertheless play a role in protein-ligand binding.

6.7 Comparison of Poisson-Boltzmann electrostatics with the enthalpies of hydration

A comparison of the hydration enthalpies of high density sites ($< 1.5\times$ bulk) with the APBS output is shown in Figure 6.7. The dataset includes 85279 sites. There is quite an interesting wide scatter in enthalpies of hydration in regions when the average magnitude of the electrostatic potential is near zero. The highly positive enthalpies of hydration in this region is likely to be due to poorly solvated on the protein surface. As the average magnitude of the electrostatic potential exceeds over $\approx 3\ k_{\mathrm{B}}Te_c^{-1}\mathrm{site}^{-1}$ all hydration enthalpies have negative signs and stabilise waters suggesting some kind of electrostatic interaction with the environment (here e_c^{-1} is the charge of an electron). However, any further correlation between the strength of the enthalpy of hydration and the magnitude of the electrostatic field is very weak. This analysis supports the hypothesis that for an accurate understanding of the stability of water molecules and their interactions, local interactions including those which are nonelectrostatic must be investigated in detail.

This is more clearly portrayed in Figure 6.8, showing three outliers found in the 1E1X protein structure of cyclin-dependent kinase 2. Figure 6.8A displays a case with low magnitude of the electrostatic potential (due possibly to shielding from bulk water) but large negative enthalpy of hydration as a result of a good coordination of an oxygen water with two arginine side chain nitrogens, an interaction with an aspartate side-chain oxygen with one of the water's hydrogens, and finally an interaction with another water molecule. Figure 6.8B shows a case of low enthalpy and high average electrostatic potential of a well buried site which coordinates well with arginine nitrogens acting as hydrogen-bond donors and an aspartate oxygen and the backbone carbonyl oxygen of a threonine acting as hydrogen-bond acceptors of a single water. Finally, Figure 6.8C shows that even if there is a large magnitude of the electrostatic potential, this does not correlate with the strength of the hydration enthalpy of the particular region. In this case the region is more buried but too poorly coordinated to allow better enthalpy of hydration. All of these cases show the importance of the local coordination environment.

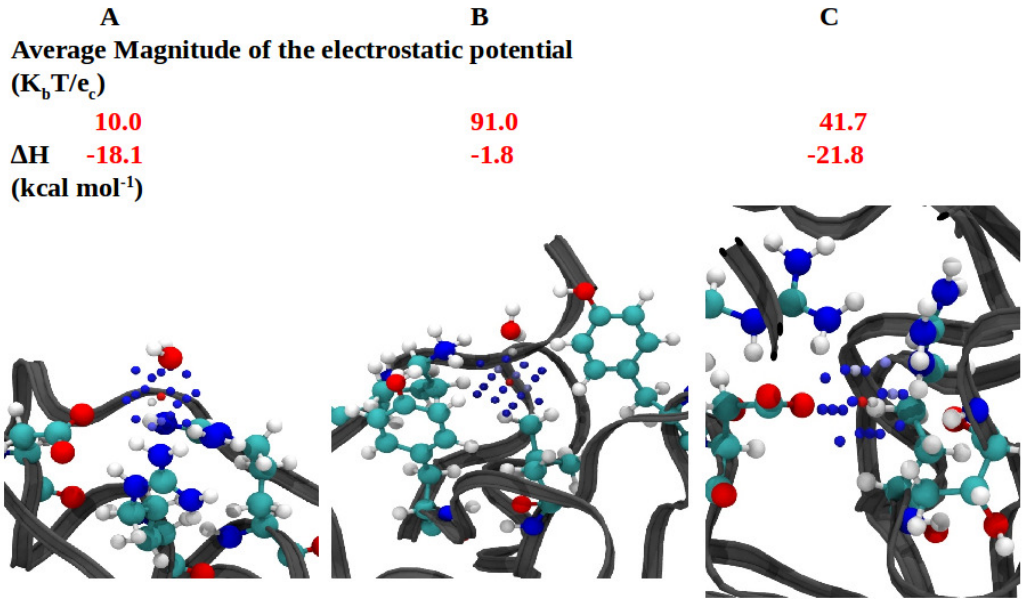


Figure 6.8: Selected outliers in the correlation of the average magnitude of the electrostatic potential (computed from APBS) and the enthalpy of hydration correlation of the PDB structure 1E1X. A) B) C) denote various cases were the electrostatics poorly correlates with a water site's free energy. Grid points related to the centroid are coloured from low relative water density to high relative water density using a colour range from blue-white-red.

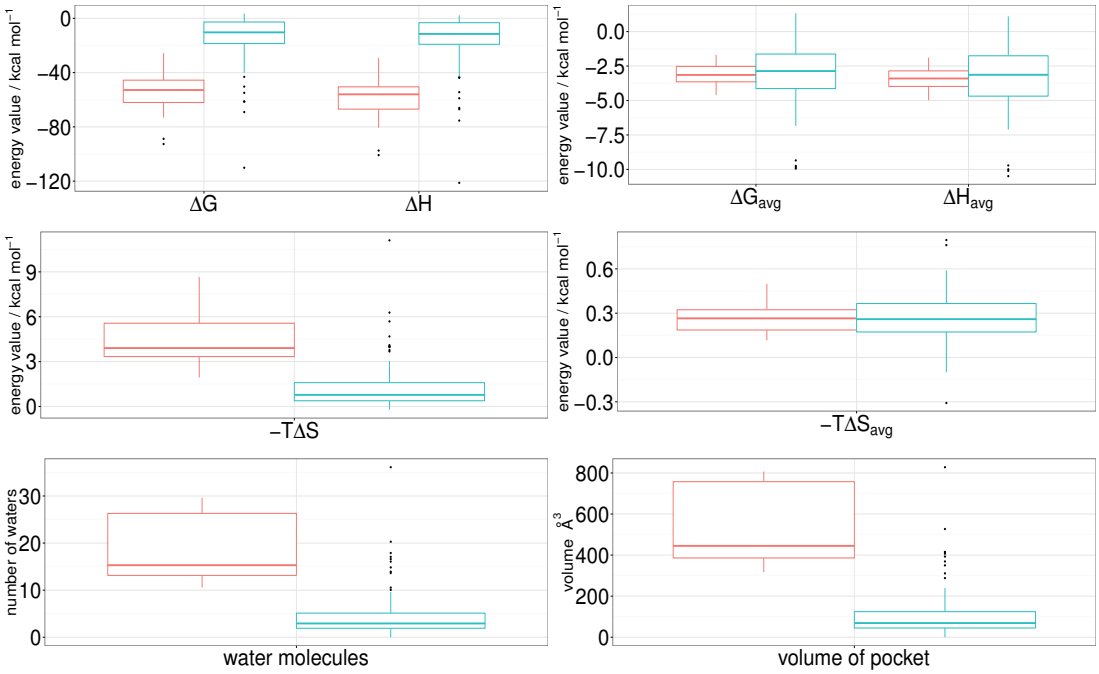


Figure 6.9: Comparison of pockets (blue) to binding sites (red) are shown as box plots which shows the median and the upper and lower quartile. The following parameter distributions for free energy, enthalpy and entropy of hydration for the entire pocket as well as the average per water value are plotted. As well as the distributions of the number of water molecules and the volumes of the pockets. In both case outliers outside of 1.5× of the interquartile range are shown.

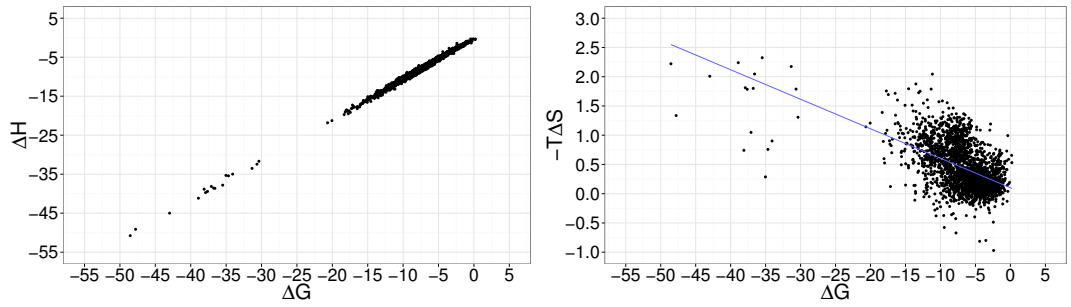


Figure 6.10: Left: plot of the free energy of hydration against the enthalpy of hydration of high density water sites at least 10× greater than bulk. Right: plot of the free energy of hydration against $-T\Delta S$ of hydration of these sites

6.7.1 Pockets compared to binding sites

From this analysis one is able to discern the global hydration properties of pockets as compared to binding sites. Pockets were located by fpocket and the free energy, enthalpy, and entropy of hydration per-water and per-site were investigated using GCT. As well as this, the volume of the site and average water number per site were also investigated (shown in Figure 6.9). First, the free energy and enthalpy of hydration are discussed. It is evident that most of the free energy of the pockets is largely enthalpic in nature in both the per-water and per-site case. There is large difference ≈ 40 kcal mol^{-1} in the median value of the free energy per site value for a pocket compared to a binding site. The difference between pockets and binding sites becomes less clear in the free energy and enthalpy per-water box plots, which have similar medians with larger variability in their distributions.

Anticorrelated trends are seen in the per-site entropy statistics, where lower entropy seems to correlate with the lower enthalpy likely due to configurational entropy loss from the stronger interactions. This is again not apparent in the per-water box plot.

Finally, the number of water molecules and volume of the pocket is investigated, and it was found that usually pockets have eight less waters than are found in a typical binding site, which is reflected in a much larger volume found in a binding site compared to a typical pockets.

This shows that binding sites globally are typically made up of more waters than an average pocket with a larger volume than normal pockets. The analysis show in Figure 6.9 suggests that volume is the most obvious indicator of a binding site and that on average water molecules in binding sites are not more or less stable than water near other pockets.

6.7.2 High density water sites

Analysis of highly clustered sites at least 10× greater than bulk density are analysed. Interestingly all these highly dense sites are all localised on the protein hydration layer. The thermodynamics of these sites are further discussed.

First, the correlation between the hydration free energy and enthalpy of these water sites is shown in Figure 6.10. A clear correlation between the enthalpy and free energy. However, there is a weak anticorrelation between the entropy and free energy. This suggests that stabilisation around the protein hydration shell tends to be driven by enthalpic stabilisation with noisy decrease in free energy of hydration created by the entropy due to the strong interactions. Outliers in the plot of the free energy of hydration against $-T\Delta S$ of hydration are visualised in Figure 6.11. In Figure 6.11A there is a very tightly bound water molecule with a free energy of -48.5 and $-T\Delta S$ of hydration of 2.22 kcal mol^{-1}. This buried water site contains two other buried waters nearby and. is coordinated by those waters as well as interacting with two hydrogen-bond acceptors. A threonine's carbonyl oxygen and also a glutamate oxygen of its carboxylate group interact with the water. The aspartate helps stabilise the other waters in the network. Due to the buried nature of the site the hydration $-T\Delta S$ increases. Looking at the site in Figure 6.11B the water is more solvent accessible and connects to bulk. The water molecule interacts with the amide side chain nitrogen atom as well as the amide backbone nitrogen of glutamine which both act as hydrogen-bond donors. The water site shows a slight density to the right where the water interacts more strongly with carbonyl oxygens of both histidine and aspartate which act as hydrogen-bond acceptors. The space in the site allows for many rotations between hydrogen bond donors and acceptors aiding the librational, and orientational entropies ($-T\Delta S_{\text{lib}}$ = -0.40,

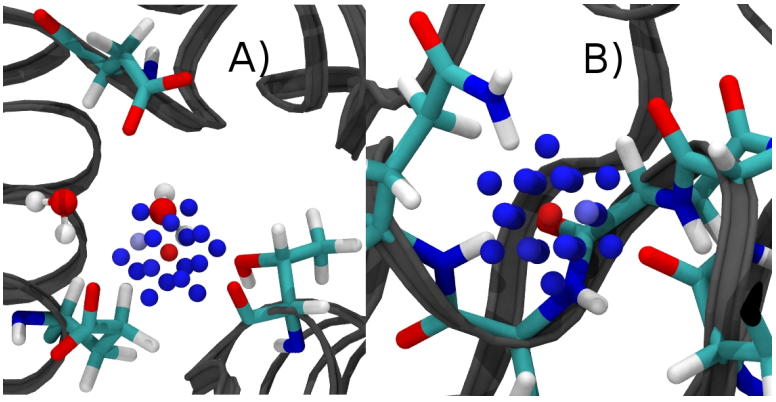


Figure 6.11: Outliers of the high density sites (10× greater than bulk) are shown. A) shows and outlier found in the 1OYN simulation with a free energy of hydration of -48.5 for the site with 2.22 $-T\Delta S$ of hydration. B) shows another case in 1E66 simulation with a free energy of hydration of -7.78 and $-T\Delta S$ of hydration of -0.67. Note, all units are in kcal mol^{-1}. Grid points related to the centroid are coloured from low density to high relative water density using a colour range from blue-white-red.

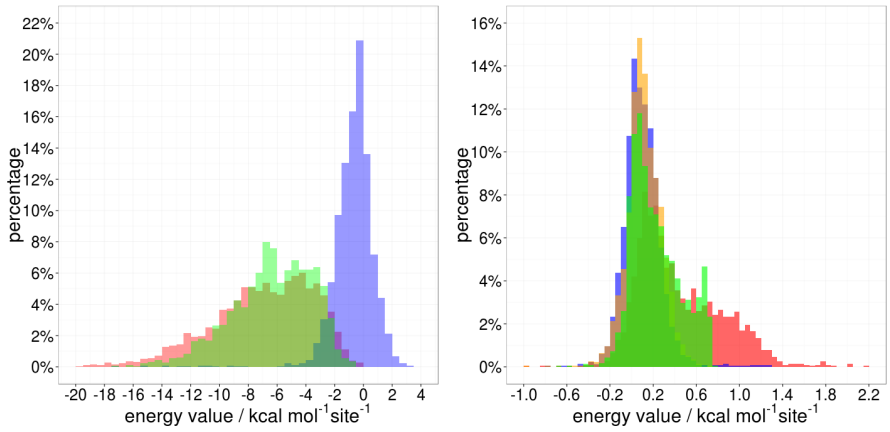


Figure 6.12: Left: probability distribution of the components of the enthalpy of hydration (red), ΔH_w (blue) and ΔH_X (green). Right: probability distribution of the components of the entropy of hydration (red), $-T\Delta S^{\text{ori}}_{w,X}$ (green) and $-T\Delta S^{\text{lib}}_{w,X}$(orange) and $-T\Delta S^{\text{vib}}_{w,X}$(blue)

$-T\Delta S_{\text{ori}} = 0.43$) while the vibrational motion is slightly restricted ($-T\Delta S_{\text{vib}} = 0.16$).

Next, the components of the hydration free energy are investigated and are shown in Figure 6.12. The distributions of the enthalpy and entropy components help analyse how they contribute to the free energy distribution of sites in the dataset. From figure 6.12(left) the enthalpy of hydration components are shown. In the red is the enthalpy of hydration is shown, blue shows the water-water enthalpies and in green is the water-solute enthalpy. From looking at the distributions one can see that the majority of the time less negative sites are composed of sites were water-water enthalpies are found while the more negative sites are dominated by water-solute enthalpy. When looking at

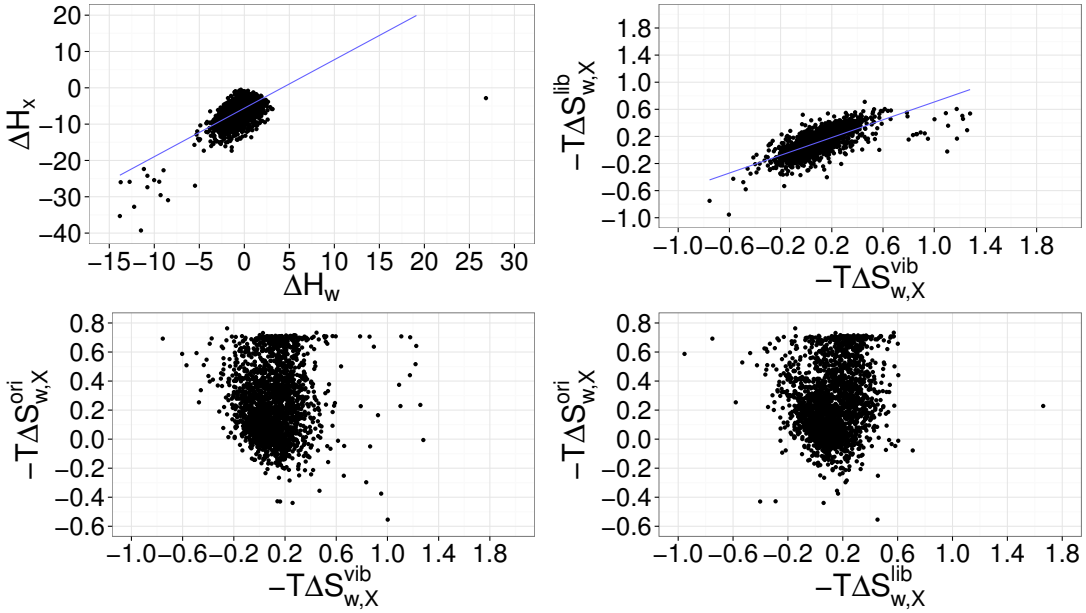


Figure 6.13: Correlation plots between the two enthalpic components and three entropic components. All values are in kcal mol^{-1}

figure 6.12(right) there is the entropy of hydration in red, the orientational entropy in green, the librational entropy in orange and vibrational in blue. From this plot it seems that all components contribute equally to lower entropy sites at higher entropy sites orientational entropy seems to have a larger contribution to the increase in hydration entropy. This shows that sites with higher entropy tends to be driven by a decrease in orientational entropy in this dataset. For the hydration enthalpy at more stabilised water sites there is usually a strong interaction with the protein ΔH_X shown in Figure 6.13. Lower stability and unstable sites are more likely to be stabilised by other waters, ΔH_w (see Figure 6.13). Finally, inspection of the entropic components reveal that larger decrease in the entropy values is caused by a loss of orientational entropy (lack of neighbours). Entropy loss in the protein hydration layer has a maximum of 2.5 kcal mol^{-1} slightly higher than the experimental limits of 2 kcal mol^{-1} suggested by Dunitz [71]. However, any entropic increase seems to be derived mostly from both the librational and vibrational entropies rather than from the orientational entropy (see Figure 6.13). Interestingly, looking at the vibrational and librational entropies in Figure 6.13, there is a good correlation between vibrational and librational entropy loss.

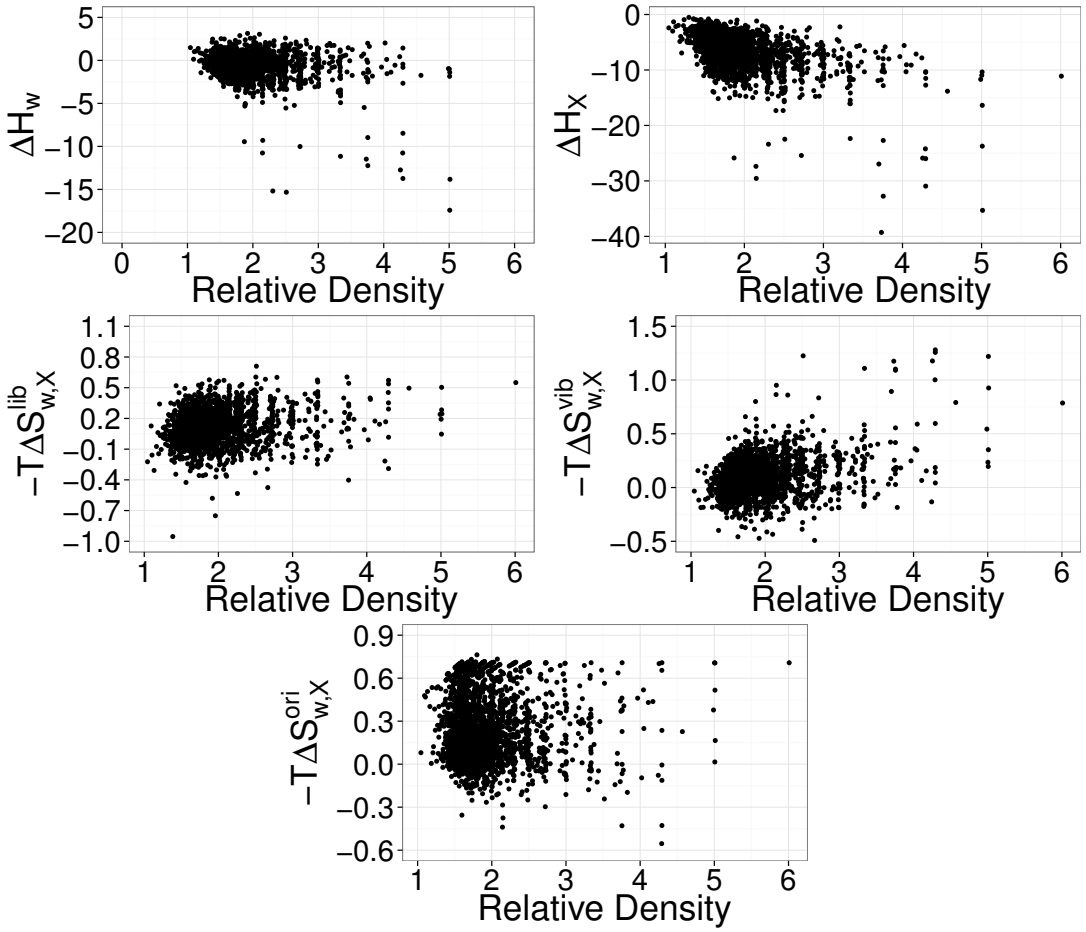


Figure 6.14: Correlation plots of the two enthalpic components and three entropic components with respect to the relative density of bulk. Enthalpy and entropy components are in kcal mol^{-1}, while the relative density is unitless.

However the scale of the loss/gain is slightly larger from the vibrational entropy. The orientational entropy does not correlate with either vibrational or librational entropies but usually contributes the most to the hydration entropy loss.

In Figure 6.14 the relationship between the various components and the relative density is examined. For the ΔH_w, there is only a weak anticorrelation between the density and the strength. The ΔH_X term shows stronger anticorrelation at relative densities less than 3. After a relative density of 3 there are few samples. Now looking at the entropy components there is no correlation with density in orientational entropy component. There is a small positive correlation with both the librational and vibrational terms but the vibrational correlation is stronger. Overall, this shows that only the strongly negative ΔH_X anticorrelates with the relative density which relates to the strong interaction localising the water. There is a weak correlation with the vibrational term probably compensating for stronger enthalpy at higher densities. All other terms do not have strong correlations with the relative density showing that high density does not necessarily entail thermodynamically stable water molecules.

6.8 Conclusions

From the large dataset of proteins analysed, an understanding of how proteins and water interact can be understood has developed. Qualitative results are similar to those found in a Watermap study by Beuming *et al.* [144] with similar ordering of amino acid chemotypes, but a lack of agreement on magnitudes of the free energy of hydration could relate to stronger filtering by density of the clusters of water, as well as a difference in the computation of the entropy. Comparisons between Poisson-Boltzmann calculations and GCT analysis indicate the need to relate water stability to local coordination environments. An analysis of binding sites compared with pockets show that binding sites tend to contain regions with more negative hydration free energies. This is because binding sites tend to be larger than other pockets. The overall hydration free energy of water sites correlates with hydration enthalpy and are anticorrelated with the hydration entropy. The high-density hydration sites are stabilised mostly by the enthalpy of interactions between the protein and water, and occasionally entropically stabilised by larger vibrational and librational modes within these. Further studies could be made to more clearly understand how hydration thermodynamics are influenced by more flexible protein environments. Work also needs to be done on finding efficient methods to characterise structural descriptors which stabilise a water molecule which could potentially be used in an analysis of the hydration free energy. Finally, it would be interesting to repeat the analysis using more elaborate definitions of the orientational entropy term such as those being developed by Henchman and coworkers in current calculations [111].

Chapter 7

Conclusions and future directions

"We shall not cease from exploration and the end of all our exploring will be to arrive where we started and know the place for the first time. Through the unknown, remembered gate when the last of earth left to discover is that which was the beginning; at the source of the longest river the voice of the hidden waterfall and the children in the apple-tree not known, because not looked for but heard, half-heard, in the stillness between two waves of the sea." - T.S. Eliot

In this chapter the research is summarised and is placed in the context of the field.

7.1 A summary

Hydration is wide reaching field with applications in all fields of chemistry. Water more and more seems play an important role in biomolecular recognition often aiding or hindering ligand binding. For this reason it is hoped that GCT can help reveal some of the energetics of these binding events. Headway has been made with validation and recent studies on protein-ligand systems accomplished.

Starting from chapter 3, GCT has been validated and reproduces experimental hydration thermodynamics of a diverse set of small molecules of varying polarities and charges, representative of those found in a biomolecular context [150]. It can help elucidate the costs of water displacement from a protein binding site [114] and water energetics in a binding process in general (chapter 5).

GCT was more thoroughly used to investigate congeneric series of Factor Xa, and HSP90a protein-ligand systems. Its use helped suggest potential drivers of the binding events. In the case of Factor Xa the best predictive value came from the protein-ligand interaction descriptor suggest that maybe solvation is less of a driver than the

protein-ligand interaction for discriminating between ligands in this series. In the case of HSP90a, just the solvation descriptors, (LIG and APO) provided good predictive values. In this case there were a few important bridging waters, and stable ligand-protein binding conformations which involves strong bridging waters, which explains why water properties correlate with binding affinities in congeneric ligands targeting HSP90a. Finally, GCT was used to evaluate how an amino acid may stabilise nearby waters. Some general trends can be discerned, but overall the local environment of each water appears to be of greater importance. In general, it is observed that high-density water sites tend to be close to crystal-water sites, but occasionally some sites do not corroborate. This could indicate resolution problems in X-ray crystallography. Lastly, water molecules in binding sites were on average just as stable as in other pockets, binding sites were better hydrated due to a larger volume accommodating a greater number of water molecules.

7.2 Strengths: GCT can aid chemical intuition

The strength of GCT is the visualisation of hydration thermodynamics from a single molecular dynamics simulation. Other methods such as FEP and TI typically have better correspondence with experiment but it is often hard to rationalise which chemical groups drive free energy differences. Visualisation and spatial decomposition of hydration thermodynamics helps understand which functional groups drive stabilisation of water molecules. This is vital for ligand optimisation and the design of new molecular entities, since different functional groups will interact with the protein and water differently. The water displacement cost can be determined from the simulations. Important water sites can then be ranked making a drug designer aware of difficult areas to fill during the ligand optimisation process. One other method capable of this kind of decomposition from a single MD simulation is the GIST method [26]. However, the binning and convergence of GCT appear faster for the entropic component due to the molecule centred harmonic-oscillator, mean-field approximations used in cell theory. However, both methods have the same weakness in the enthalpy prediction which relies on the force field used as well as sufficient sampling.

As well as for its practical uses GCT can also be used to investigate the nature of enthalpy/ entropy compensation during the binding process in future studies. For instance the dewetting process can be analysed further in terms of entropic components found in GCT could help have a better understanding of the binding process. It can be envisaged that different GCT calculation can be applied at various steps of a reaction coordinate between the bound and unbound states.

For these reasons, it is hoped that future use of GCT could further elucidated the role

of solvent in the biomolecular context.

7.3 Weaknesses and new possible directions

The major issue with all GCT approaches is the lack of a treatment of the entropy of the protein and ligand. It is thought that computing the macromolecular entropy of both the protein and ligand from forces using a method described by Hensen *et al.* [130], can provide a complete description of the entire binding process. Such a simulation would then require conformation-sensitive grids which may recognise different conformational states and cluster different snapshots on-the-fly or post-simulation. Unfortunately, such simulations would require proper sampling of a sufficient number of distinct conformational clusters. One option would be to process conformations separately and weight their relative contributions to the binding process. Similar flexible methods have been attempted by researchers in Novartis with time-averaged charges, and hydrogen, and oxygen densities of the water molecules as well as relevant solute atoms which could be hydrogen-bonding. Analysis is aided with anchoring restraints which makes analysis easier [151].

GCT could also be used in the context of QM/MM or *ab initio* MD simulations to estimate entropies from quantum forces/torques. This may be more pertinent for more highly charged systems and others containing metals frequently found in certain enzymes. This could help understand the role of water in enzymatic catalysis which frequently can occur.

7.4 Conclusion

Overall, GCT could become a practical tool for CADD if some practical improvements in speedup could occur. It could be used to assess protein binding sites and aid ligand optimisation by assessing water displacement costs. It can also give a good estimate of the enthalpy and entropy components of hydration thermodynamics. Care needs to be taken to be sure that correct conformational states are being examined for both the protein and the ligand. The tool could also be used to help assess the stability of waters identified through NMR, and X-ray crystallography. It can also assess small molecule free energy hydration in general and identify where negative and positive contributions are made. It could inform different fields of chemistry including organic synthesis, materials science, and indeed any molecular study where water is involved. However, in those cases appropriate parameterisation of water interactions to the particular atoms of interest would have to be available assuming that the system contains strong interactions between molecules where the harmonic approximation of GCT work best. In

conclusion, GCT was successfully used for biomolecular simulations to gain insights into protein-ligand binding in the context of drug design.

Bibliography

[1] Xu, D.; Williamson, M. J.; Walker, R. C. Chapter 1 - Advancements in Molecular Dynamics Simulations of Biomolecules on Graphical Processing Units. In , Vol. 6; Wheeler, R. A., Ed.; Elsevier: 2010.

[2] Michel, J.; Foloppe, N.; Essex, J. W. *Mol Inform* **2010,** *29,* 570-578.

[3] DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. *J Health Econ* **2003,** *22,* 151-185.

[4] Michel, J. *Phys Chem Chem Phys* **2014,** *16,* 4465-4477.

[5] Kapetanovic, I. *Chem-Biol Interact* **2008,** *171,* 165-176.

[6] Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. *J Chem Theory Comput* **2009,** *5,* 350-358.

[7] Lam, P. Y.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N. *Science* **1994,** *263,* 380–384.

[8] Ladbury, J. E. *Chem Biol* **1996,** *3,* 973–980.

[9] Ball, P. *Nature* **2008,** *452,* 291-292.

[10] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J Chem Phys* **1983,** *79,* 926-935.

[11] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *Intermolecular Forces;* D. Reidel Publishing Company: 1981.

[12] Chaplin, M. "http://www1.lsbu.ac.uk/water/water_models.html", .

[13] Wickstrom, L.; Okur, A.; Simmerling, C. *Biophys J* **2009,** *97,* 853-856.

[14] Hess, B.; van der Vegt, N. F. A. *J Phys Chem B* **2006,** *110,* 17616-17626.

[15] Giuffre, E.; Prestipino, S.; Saija, F.; Saitta, M.; Giaquinta, P. *J Chem Theory Comput* **2010,** *6,* 625-636.

[16] Fenley, A. T.; Henriksen, N. M.; Muddana, H. S.; Gilson, M. K. *J Chem Theory Comput* **2014,** *10,* 4069-4078.

[17] Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J Phys Chem B* **2007,** *111,* 2242-2254.

[18] Joung, I. S.; Cheatham, Thomas E., I. *J Phys Chem B* **2008,** *112,* 9020-9041.

[19] Woodhead, A. *et al. J Med Chem* **2010,** *53,* 5956-5969.

[20] Adler, M.; Davey, D. D.; Phillips, G. B.; Kim, S. H.; Jancarik, J.; Rumennik, G.; Light, D. R.; Whitlow, M. *Biochemistry* **2000,** *39,* 12534-12542.

[21] Huang, N.; Shoichet, B. K. *J Med Chem* **2008,** *51,* 4862-4865.

[22] Li, Z.; Lazaridis, T. *J Phys Chem B* **2004,** *109,* 662-670.

[23] Breiten, B.; Lockett, M. R.; Sherman, W.; Fujita, S.; Al-Sayah, M.; Lange, H.; Bowers, C. M.; Heroux, A.; Krilov, G.; Whitesides, G. M. *J Am Chem Soc* **2013,** *135,* 15579-15584.

[24] Tamura, A.; Privalov, P. L. *J Mol Biol* **1997,** *273,* 1048-1060.

[25] Allen, F. H. *Acta Crystallogr B* **2002,** *58,* 380-388.

[26] Nguyen, C.; Young, T. K.; Gilson, M. *J Chem Phys* **2012,** *137,* 044101.

[27] Young, T.; Hua, L.; Huang, X.; Abel, R.; Friesner, R.; Berne, B. J. *Proteins: Struct, Funct, Bioinf* **2010,** *78,* 1856-1869.

[28] Lazaridis, T. *J Phys Chem B* **2000,** *104,* 4964-4979.

[29] Lazaridis, T. *J Phys Chem B* **1998,** *102,* 3542-3550.

[30] Lazaridis, T. *J Phys Chem B* **1998,** *102,* 3531-3541.

[31] Irudayam, S.; Henchman, R. *J Phys Chem B* **2009,** *113,* 5871-5884.

[32] Siebert, X.; Amzel, L. M. *Proteins: Struct, Funct, Bioinf* **2004,** *54,* 104-115.

[33] Hummer, G. *Nat Chem* **2010,** *2,* 906-907.

[34] Moghaddam, S.; Yang, C.; Rekharsky, M.; Ko, Y. H.; Kim, K.; Inoue, Y.; Gilson, M. *J Am Chem Soc* **2011,** *133,* 3570-3581.

[35] Haider, K.; Huggins, D. *J Chem Inf Model* **2013,** *53,* 2571-2586.

[36] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J Comput Chem* **1983,** *4,* 187-217.

[37] Berendsen, H.; van der Spoel, D.; van Drunen, R. *Comput Phys Commun* **1995,** *91,* 43-56.

[38] Case, D.; Cheatham, T.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. *J Comput Chem* **2005,** *26,* 1668-1688.

[39] Beauchamp, K. A.; Lin, Y.-S.; Das, R.; Pande, V. S. *J Chem Theory Comput* **2012,** *8,* 1409-1414.

[40] Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins* **2010,** *78,* 1950-1958.

[41] Nerenberg, P. S.; Head-Gordon, T. *J Chem Theory Comput* **2011,** *7,* 1220-1230.

[42] Li, D.-W.; Brüschweiler, R. *Angew Chem* **2010,** *122,* 6930–6932.

[43] Best, R. B.; Hummer, G. *J Phys Chem B* **2009,** *113,* 9004-9015.

[44] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J Comput Chem* **2004,** *25,* 1157-1174.

[45] Jakalian, A.; Jack, D.; Bayly, C. *J Comput Chem* **2002,** *23,* 1623-1641.

[46] Shivakumar, D.; Deng, Y.; Roux, B. *J Chem Theory Comput* **2009,** *5,* 919-930.

[47] Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J Chem Phys* **1982,** *76,* 637-649.

[48] Andersen, H. *J Chem Phys* **1980,** *72,* 2384-2393.

[49] Uberuaga, B.; Anghel, M.; Voter, A. *J Chem Phys* **2004,** *120,* 6363-6374.

[50] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J Chem Phys* **1984,** *81,* 3684-3690.

[51] Tironi, I.; Sperb, R.; Smith, P.; van Gunsteren, W. *J Chem Phys* **1995,** *102,* 5451-5459.

[52] Leach, A. *Molecular Modelling: Principles and Applications (2nd Edition);* Prentice Hall: 2 ed.; 2001.

[53] Nguyen, C. N.; Cruz, A.; Gilson, M. K.; Kurtzman, T. *J Chem Theory Comput* **2014,** *10,* 2769-2780.

[54] Goodford, P. J. *J Med Chem* **1985,** *28,* 849-857.

[55] Grant, J. A.; Pickup, B. T.; Nicholls, A. *J Comput Chem* **2001,** *22,* 608-640.

[56] Zhou, S.; Cheng, L.-T.; Dzubiella, J.; Li, B.; McCammon, J. A. *J Chem Theory Comput* **2014,** *10,* 1454-1467.

[57] Zhou, S.; Rogers, K. E.; de Oliveira, C. A. F.; Baron, R.; Cheng, L.-T.; Dzubiella, J.; Li, B.; McCammon, J. A. *J Chem Theory Comput* **2013,** *9,* 4195-4204.

[58] Truchon, J.-F.; Pettitt, B. M.; Labute, P. *J Chem Theory Comput* **2014,** *10,* 934-941.

[59] Michel, J.; Tirado-Rives, J.; Jorgensen, W. *J Phys Chem B* **2009,** *113,* 13337-13346.

[60] Fedorov, D. G.; Kitaura, K. *J Phys Chem A* **2007,** *111,* 6904-6914.

[61] Henchman, R. *J Chem Phys* **2007,** *126,* 064504.

[62] Irudayam, S.; Plumb, R.; Henchman, R. *Faraday Discuss* **2010,** *145,* 467-485.

[63] Irudayam, S. J.; Henchman, R. H. *J Phys: Condens Matter* **2010,** *22,* 284108.

[64] Irudayam, S. J.; Henchman, R. H. *Mol Phys* **2011,** *109,* 37-48.

[65] Nicholls, A.; Mobley, D.; Guthrie, P.; Chodera, J.; Bayly, C.; Cooper, M.; Pande, V. *J Med Chem* **2008,** *51,* 769-779.

[66] "SZMAP Theory", "https://docs.eyesopen.com/szmap/theory.html".

[67] Truchon, J. F.; Nicholls, A.; Grant, J. A.; Iftimie, R. I.; Roux, B.; Bayly, C. I. *J Comput Chem* **2010,** *31,* 811-824.

[68] Sindhikara, D. J.; Hirata, F. *J Phys Chem B* **2013,** *117,* 6718-6723.

[69] Kong, X.; Brooks, C. *J Chem Phys* **1996,** *105,* 2414-2423.

[70] Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys J* **1997,** *72,* 1047-1069.

[71] Dunitz, J. D. *Science (New York, N.Y.)* **1994,** *264,* 670-670.

[72] Nguyen, C.; Gilson, M. K.; Young, T. *ArXiv e-prints* **2011,** "http://arxiv.org/abs/1108.4876".

[73] Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids (Oxford Science Publications);* Oxford science publications Oxford University Press: Reprint ed.; 1989.

[74] Mie, G. *Annalen der Physik* **1903,** *316,* 657-697.

[75] Henderson, D. *Annu Rev Phys Chem* **1964,** *15,* 31-62.

[76] Eisenstein, A.; Gingrich, N. S. *Phys Rev* **1942,** *62,* 261-270.

[77] Lennard-Jones, J. E.; Devonshire, A. F. *Proc R Soc A* **1937,** *163,* 53-70.

[78] Henchman, R. H. *J Chem Phys* **2003,** *119,* 400-406.

[79] Karplus, M.; Kushick, J. N. *Macromolecules* **1981,** *14,* 325-332.

[80] Gerogiokas, G.; Calabro, G.; Henchman, R. H.; Southey, M. W. Y.; Law, R. J.; Michel, J. *J Chem Theory Comput* **2014,** *10,* 35-48.

[81] Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J Chem Phys* **2004,** *120,* 9665-9678.

[82] Bondi, A. *J Phys Chem* **1964,** *68,* 441-451.

[83] Guillot, B.; Guissani, Y. *J Chem Phys* **1993,** *99,* 8075-8094.

[84] Case, D. *et al. Amber 11;* **2010**.

[85] Miyamoto, S.; Kollman, P. *J Comput Chem* **1992,** *13,* 952-962.

[86] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. *J Comput Phys* **1977,** *23,* 327-341.

[87] Darden, T.; York, D.; Pedersen, L. *J Chem Phys* **1993,** *98,* 10089-10092.

[88] Sagui, C.; Darden, T. *Simulation and Theory of Electrostatic Interactions in Solution;* Melville, NY, 1999.

[89] Woods, C.; Michel, J. "Sire Molecular Simulation Framework, Revision 1786", 2013.

[90] Eastman, P. *et al. J Chem Theory Comput* **2012,** *9,* 461-469.

[91] Jorgensen, W.; Ravimohan, C. *J Chem Phys* **1985,** *83,* 3050-3054.

[92] Zwanzig, R. *J Chem Phys* **1954,** *22,* 1420-1426.

[93] Shyu, C.; Ytreberg, M. *J Comput Chem* **2009,** *30,* 2297-2304.

[94] Beutler, T.; Mark, A.; van Schaik, R.; Gerber, P.; van Gunsteren, W. *Chem Phys Lett* **1994,** *222,* 529-539.

[95] Zacharias, M.; Straatsma, T. P.; McCammon, J. A. *J Chem Phys* **1994,** *100,* 9025-9031.

[96] Michel, J.; Verdonk, M.; Essex, J. *J Chem Theory Comput* **2007,** *3,* 1645-1655.

[97] Wagner, W.; Pruss, A. *J Phys Chem Ref Data* **2002,** *31,* 387-535.

[98] Irudayam, S. J.; Henchman, R. H. *J Chem Phys* **2012,** *137.*

[99] Shelton, D. P. *J Chem Phys* **2012,** *136,* 044503.

[100] Huggins, D.; Payne, M. *J Phys Chem B* **2013,** *117,* 8232-8244.

[101] Michel, J.; Orsi, M.; Essex, J. W. *J Phys Chem B* **2008,** *112,* 657-660.

[102] Schmid, R.; Miah, A.; Sapunov, V. *Phys Chem Chem Phys* **2000,** *2,* 97-102.

[103] Kastenholz, M. A.; Hunenberger, P. H. *J Chem Phys* **2006,** *124.*

[104] Peter, C.; Oostenbrink, C.; van Dorp, A.; van Gunsteren, W. F. *J Chem Phys* **2004,** *120,* 2652-2661.

[105] Kastenholz, M. A.; Hunenberger, P. H. *J Chem Phys* **2006,** *124.*

[106] Reif, M. M.; Hunenberger, P. H. *J Chem Phys* **2011,** *134.*

[107] Naim, B.; Marcus, Y. *J Chem Phys* **1984,** *81,* 2016-2027.

[108] Gatta,; Barone, G.; Elia, V. *J Solution Chem* **1986,** *15,* 157-167.

[109] Wolfenden, R. *Biochemistry* **1978,** *17,* 201-204.

[110] Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J Solution Chem* **1981,** *10,* 563-595.

[111] Henchman, R.; Cockram, S. *Faraday Discuss* **2013**.

[112] Huggins, D. *Phys Chem Chem Phys* **2012,** *14,* 15106-15117.

[113] Henchman, R. H.; McCammon, J. A. *J Comput Chem* **2002,** *23,* 861-869.

[114] Gerogiokas, G.; Southey, M.; Mazanetz, M.; Heifetz, A.; Bodkin, M.; Law, R.; Michel, J. *Phys Chem Chem Phys* **2015,** 8416-8426.

[115] Alvarez-Garcia, D.; Barril, X. *J Chem Theory Comput* **2014,** *10,* 2608-2614.

[116] Lexa, K. W.; Carlson, H. A. *J Chem Inf Model* **2013,** *53,* 391-402.

[117] Yu, W.; Lakkaraju, S. K.; Raman, E. P.; MacKerell, A. D., J. *J Comput-Aided Mol Des* **2014,** *28,* 491-507.

[118] Bodnarchuk, M. S.; Viner, R.; Michel, J.; Essex, J. W. *J Chem Inf Model* **2014,** *54,* 1623-1633.

[119] Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. *Proc Natl Acad Sci U.S.A.* **2007,** *104,* 808-813.

[120] Brodney, M. A. *et al. J Med Chem* **2012,** *55,* 9224-9239.

[121] Chen, J. M.; Xu, S. L.; Wawrzak, Z.; Basarab, G. S.; Jordan, D. B. *Biochemistry* **1998,** *37,* 17735-17744.

[122] Liu, C. *et al. J Med Chem* **2005,** *48,* 6261-6270.

[123] Wissner, A. *et al. J Med Chem* **2000,** *43,* 3244-3256.

[124] Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. *J Am Chem Soc* **2009,** *131,* 15403-15411.

[125] Trott, O.; Olson, A. J. *J Comput Chem* **2010,** *31,* 455-461.

[126] Seeliger, D.; de Groot, B. L. *J Comput-Aided Mol Des* **2010,** *24,* 417-422.

[127] Loeffler, H. "FESetup", "https://ccpforge.cse.rl.ac.uk/gf/project/ccpbiosim/", 2014.

[128] Alvarez-Garcia, D.; Barril, X. *J Med Chem* **2014,** *57,* 8530-8539.

[129] Olano, R.; Rick, S. *J Am Chem Soc* **2004,** *126,* 7991-8000.

[130] Hensen, U.; Gräter, F.; Henchman, R. H. *J Chem Theory Comput* **2014,** *10,* 4777-4781.

[131] Michel, J.; Essex, J. W. *J Comput-Aided Mol Des* **2010,** *24,* 639-658.

[132] Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. *J Am Chem Soc* **2008,** *130,* 2817-2831.

[133] Taha, M. O.; Qandil, A. M.; Zaki, D. D.; AlDamen, M. A. *Eur J Med Chem* **2005,** *40,* 701-727.

[134] Murray, C. W. *et al. J Med Chem* **2010,** *53,* 5942-5955.

[135] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct, Funct, Bioinf* **2006,** *65,* 712-725.

[136] "Chemicalize", "http://www.chemicalize.org/", 2015.

[137] Pearlman, D.; Charifson, P. *J Med Chem* **2001,** *44,* 502-511.

[138] Perzborn, E.; Roehrig, S.; Straub, A.; Kubitza, D.; Misselwitz, F. *Nat Rev Drug Discov* **2011,** *10,* 61-75.

[139] England, J. L.; Pande, V. S. *Biochem Cell Biol* **2010,** *88,* 359-369.

[140] Nicholls, A.; Sharp, K. A.; Honig, B. *Proteins* **1991,** *11,* 281-296.

[141] Reddy, C.; Das, A.; Jayaram, B. *J Mol Biol* **2001,** *314,* 619-632.

[142] Conte, L. L.; Chothia, C.; Janin, J. *J Mol Biol* **1999,** *285,* 2177-2198.

[143] Lu, Y.; Wang, R.; Yang, C.-Y.; Wang, S. *J Chem Inf Model* **2007,** *47,* 668-675.

[144] Beuming, T.; Che, Y.; Abel, R.; Kim, B.; Shanmugasundaram, V.; Sherman, W. *Proteins: Struct, Funct, Bioinf,* **2012,** *80,* 871-883.

[145] Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc Natl Acad Sci U.S.A.* **2001,** *98,* 10037-10041.

[146] Honig, B.; Nicholls, A. *Science* **1995,** *268,* 1144-1149.

[147] Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Jr., E. F. M.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *Arch Biochem Biophys* **1978,** *185,* 584-591.

[148] Guilloux, V. L.; Schmidtke, P.; Tuffery, P. *BMC Bioinformatics* **2009,** *10,* 168.

[149] Liang, J.; Woodward, C.; Edelsbrunner, H. *Protein Sci* **1998,** *7,* 1884-1897.

[150] Gerogiokas, G.; Henchman, R.; Southey, M.; Law, R.; Michel, J. *Abstr Pap Am Chem S* **2013,** *246.*

[151] Velez-Vega, C.; McKay, D. J. J.; Aravamuthan, V.; Pearlstein, R.; Duca, J. S. *J Chem Inf Model* **2014,** *54,* 3344-3361.

[152] Roe, D. R.; Cheatham, T. E. *J Chem Theory Comput* **2013,** *9,* 3084-3095.

Appendices

Appendix A

Nautilus workflow

Grid cell theory has been implemented into a python code called *Nautilus*. In this section workflows are described to indicate how excess free energies of any particular system of interest in the NPT ensemble could be computed from a molecular dynamics trajectory. In each workflow it is assumed that the simulation is centred on the solute (typically centre of mass) and the rigid body motions have been removed so that the grid obtains all relevant configurations in one frame of reference for the system.

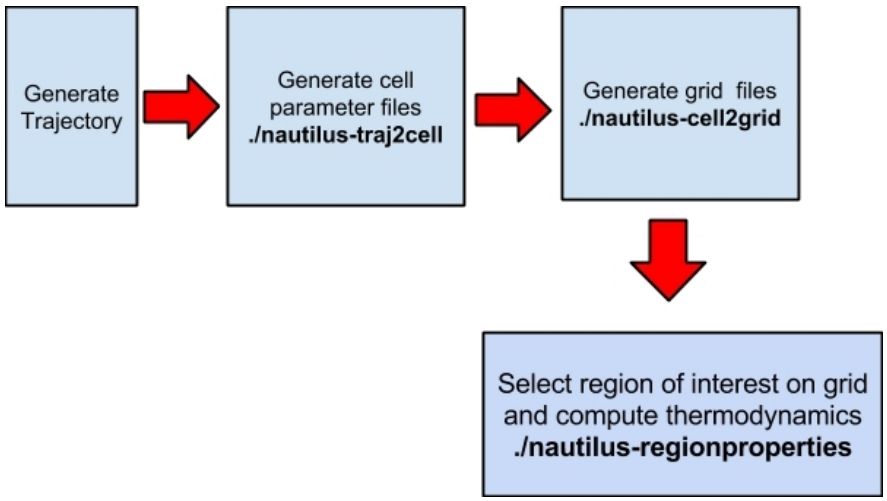


Figure A.1: The workflow for computing the excess hydration free energy of any region of space in either a ligand, protein or complex simulation. The final step where a region is selected is particularly important. Either entire regions around the ligand can be chosen or clustered regions of interest can be compared.

Each computations follows the same workflow described as figure A.1. Initially a simulation is run for any solute, the ligand, the protein or the complex. Often the protein configuration found when the ligand is bound rather than a relaxed protein structure in the solvent is used referred to as the PSAPO structure. Another alternative is the structure of a protein-ligand complex called the HOLO structure or the ligand simu-

lation alone the LIG structure. Typically these structures are derived from a X-ray crystallography or NMR structures. In the simulations either a TIP4P-EW or TIP3P water model is supported but only TIP4P-EW has been tested extensively. Any protein force field may be accommodated but only combinations of TIP4P-EW with ff99SB and TIP4P-EW with ff12SB have been tested within the work here. The following shows the entire workflow from the initial trajectory to the final desired output, an analysed grid.

Step 1: Model PSAPO/HOLO/LIG

- Produce equilibrated AMBER top/crd files for the system of interest which typically include PSAPO (heavy atom restrained protein alone found in a complex of interest).

- In $BASE and $SIM, create an input folder and move top/crd files there. Let's assume you called them SYSTEM.top and SYSTEM.crd. $SIM indicates the simulation which may be either HOLO, PSAPO, or LIG which are the required types of simulations.

mkdir $BASE/$SIM/input
mv < *where_the_file_is* > *.top* input/
mv < *where_the_file_is .crd* input/

Step 2: Run the PSAPO/HOLO/LIGAPO MD simulation

- Sire/OpenMM is used here to generate a MD trajectory but alternative MD packages can be used. In the $BASE folder create a subfolder called 'run'

- Only CHARMM DCD files are accepted as trajectory inputs. However, CPPTRAJ [152] can be used to convert and centre (to remove rigid body translation and rotation) from other trajectory formats.

mkdir $BASE/$SIM/run
cd run
ln -s ../input/SYSTEM.top
ln -s ../input/SYSTEM.crd
sommd(sire MD command or other run in another package)

Step 3: Generate intermediate cell parameter files

- If there are multiple dcd files, concatenate them so the entire trajectory is sampled.

```
cd $BASE/$SIM
mkdir analysis
cd analysis
ln -s ../run/traj00001.dcd
ln -s ../input/SYSTEM.top
ln -s ../input/SYSTEM.crd
```

- create a *nautilus* configuration file (example shown)

```
nano nautilus.cfg
# The parameters below define the grid position. If absent, the default behavior is to
use a grid that covers the complete simulation box
#
grid_center_x = 14.4  #the center of the grid
grid_center_y = 14.3
grid_center_z = 14.5
grid_plus_x = 7.0  #the positive grid extent along x
grid_min_x = 7.0  #the negative grid extent along x
grid_plus_y = 7.0
grid_min_y = 7.0
grid_plus_z = 7.0
grid_min_z = 7.0
# # The parameters below can also be passed from the command line (and will be overriden by
command line arguments)
#
#topfile =“SYSTEM.top” # default is system.top
#crdfile = “SYSTEM.crd” # default is system.crd
#trajfile = “traj000000001.dcd” # default is traj000000001.dcd
#start_frame = 0 # default is 0
#end_frame = 1000000000 # default is 1000000000
#
# The parameters below are at their default values and should not be changed/activated
# unless you are an expert
#cutoff = 10.0*angstrom # the non-bonded cutoff for intermolecular energy evaluation in
Angstrom. default is 10 angstrom.
#rfdielectric = 78.3 # the dielectric constant for the reaction field. Default is 78.3 representing
the dielectric of water.
#water_model = TIP4PEW-SireOpenMM # the CT parameterisation to use for all calculations.
Choices (TIP4PEW-SireOpenMM(default), TIP3P and TIP4PEW-RH).
```

#cell_dir = “cell” # the name of the output folder where the cell files are stored (default is ‘cell’)
frequencyupdate = 1 # how often the snapshots are updated for the running average of properties

- The configuration file defines the parameters which can be altered. The first parameters define the box centre and the distance from the centre in all directions.
- More precise parameters on how electrostatics are calculated, water models, overwrite options and more can be altered.
- If one wants to generate cell files on multiple processors, “chunked” trajectories with various start and end points can be generated with BASH scripting.
- To begin the step run the following command should be executed:

/$HOME/sire.app/bin/nautilus-traj2cell

- With the appropriate configuration file or supply arguments. -h provides a description of all the command line arguments which overwrite the configuration file parameters shown below.

usage: nautilus-traj2cell [-h] [-C [CONFIG]] [−−author] [−−version] [-t [TOPOLOGY_FILE]] [-c [COORDINATE_FILE]] [-d [DATA_FILE]] [-s [START_FRAME]] [-e [END_FRAME]] [-b]

Generate cell files from a passed trajectory

optional arguments:
-h, –help show this help message and exit
-C [CONFIG], −−config [CONFIG]
Supply an optional Nautilus CONFIG file to control the
calculation.
–author Get information about the authors of this script.
–version Get version information about this script.
-t [TOPOLOGY_FILE], −−topology_file [TOPOLOGY_FILE]
The Amber topology file containing the system.
-c [COORDINATE_FILE], −−coordinate_file [COORDINATE_FILE]
The Amber coordinate file giving the coordinates of
all of the atoms in the passed topology file.
-d [DATA_FILE], −−data_file [DATA_FILE]
The simulation trajectory file containing coordinates and box information.
-s [START_FRAME], −−start_frame [START_FRAME]
The frame number of the first frame to analyse.
-e [END_FRAME], −−end_frame [END_FRAME]

The frame number of the last frame to analyse.

-b, –benchmark Benchmark the Nautilus subroutines.

- If comparing HOLO or APO simulations the protein should have been aligned properly and rigid body translations should have been removed using the "center" command in cpptraj for instance.

Step 4: Generate grid files

- Run with arguments or with a configuration file (same format as previously shown): /$HOME/sire.app/bin/nautilus-cell2grid

Output of help command: usage: nautilus-cell2grid [-h] [-C [CONFIG]] [−−author] [−−version] [-c [CELL_DIR]] [-s [START_FRAME]] [-e [END_FRAME]] [-b]

Generate grid files from cell files containing water parameters over a whole trajectory in a particular volume defined by a grid

optional arguments:

-h, −−help show this help message and exit

-C [CONFIG], −−config [CONFIG] Supply an optional Nautilus CONFIG file to control the calculation.

−−author Get information about the authors of this script.

−−version Get version information about this script.

-c [CELL_DIR], −−cell_dir [CELL_DIR]

The Amber topology file containing the system.

-s [START_FRAME], −−start_frame [START_FRAME]

The frame number of the first frame to analyse.

-e [END_FRAME], −−end_frame [END_FRAME]

The frame number of the last frame to analyse.

-b, −−benchmark Benchmark the Nautilus subroutines.

- Edit the *nautilus* configuration file to average different parts of the trajectory.

- The start and end frame define the interval of interest. Grid area is defined the same way as for the "traj2cell" command and then the step size and grid_count_cutoff (low count grids can be removed) can be modified as well. Frequency update controls how often the running average of the thermodynamic parameters is updated.

```
# The input folder where the cell files are stored
cell_dir = "./cell"
# The ouput folder where the grid files will be saved
grid_dir = "./grid"
# The grid will be averaged using all the cell files in the time interval below. Set to -1 and 1e10
if you want to use all data. Note the interval is in frame numbers
cell_interval = (1000, 22000)
# The grid.in file defines the grid location
grid_infile = "./grid.in"
# The grid step defines the grid density in x/y/z
grid_step = 1 # Angstroms
# Grid points with less than count will be discarded
grid_count_cutoff = 0 # set at zero to make sure grids are identical
# temperature
temperature = 298 # kelvin
# averageupdate frequency
frequencyupdate = 100
# Whether to overwrite existing output
Overwrite = True
# The water model used for the simulations
waterModel = "TIP4PEW-SireOpenMM"
```

- The results are placed in the subfolder grid. There will be pdb, MOE (molecular operating environment) and dx files for every component of the excess free energy of water. The grid parameters (coordinates, forces, torques, orientational number, density, volume, energy) are in the file "grid.forces". There are also water density normalised thermodynamic files included as well.

Step 5: Computing $\Delta G^s_w(P)$, $\Delta G^s_w(PL)$, and $\Delta G^s_w(L)$ from PSAPO, HOLO, LIG simulations respectively.

- It is incorrect to sum the entropies of grid points belonging to a given region to obtain the entropy of the region because the entropy is calculated from the average forces, and torques. Instead, the grid parameters in the file "grid.forces" have to be averaged.

- To do this a region file must be created. The command "cell2grid" writes by default in the "grid" folder a region file called "all.region" that covers all the grid points.

- Regions can be created simply by writing a region file with all the grid point indices of interest. For instance, the binding site region could be selected as "bindingsite.region" file, either manually or with the help of a script. The contents of the file should look like this:

  123 124 125 546 (..)

- Where the numbers are the indices of the grid points in the binding site. The grid.forces file is found within the "grid" folder and contains all parameters for each grid point which is preceded by a grid index.

- With the appropriate region file the "nautilus-regionproperties" command can be run to obtain correct thermodynamics for the region selected.

Run the nautilus-regionproperties command:

usage: nautilus [-h] [−−author] [−−version] [-g [GRIDFORCES]] [-r [REGIONFILE]] [-b]

Generate cell files from a passed trajectory

optional arguments: -h, −−help show this help message and exit
−−author Get information about the authors of this script.
−−version Get version information about this script.
-g [GRIDFORCES], −−gridforces [GRIDFORCES]
Grid.forces file which specifies average parameters of each grid point.
-r [REGIONFILE], −−regionfile [REGIONFILE]
Region file which specifies grid points to be averaged.
-b, −−benchmark Benchmark the Nautilus subroutines.

This results in the hydration free energy of the system of interest, or more accurately of the solute conformation selected. This can be repeated for the remaining systems of interest.

Steps **6** and **7** can be followed only for studies of scoring or binding of ligands to proteins so is not necessary for small molecule hydration studies.

Step 6: Computing ΔE from HOLO simulation

- Extract interaction energies from the HOLO simulation by extracting the ligand-protein interactions directly from the simulation either from Sire/OpenMM or with CPPtraj.

Step 7: Add all the terms

FINAL SCORE=$\Delta G^s_w(PL) + \Delta E - \Delta G^s_w(P) - \Delta G^s_w(L)$of eq (1.1) on pg. 6.

ΔE = interaction energy from the HOLO simulation
$\Delta G^s_w(PL)$ = protein-ligand solvation free energy (at 4 ang from VDW or from clusters of interest)
$\Delta G^s_w(P)$ = protein solvation free energy (at 4 ang from VDW or from clusters of interest)
$\Delta G^s_w(L)$= ligand solvation free energy (at 4 ang from VDW or from clusters of interest)
Note: in parentheses are the recommended cutoffs but these can be varied. Also to reiterate all proteins and ligands are restrained in the same configuration in this protocol.

Density clustering algorithm:

This algorithm clusters grid points into centroids using densities from the grid and distance cutoffs for neighbour lists to define the size of the cluster. This is useful for focusing the analysis on high density water sites and ranking waters of interest in the binding site.

nautilus-clustergrids

usage: nautilus [-h] [–author] [−−version] [-C [CONFIG]] [-g [GRIDFORCES]] [-n [NEIGHCUT]] [-lt [LOWT]] [-b]

Cluster grid points into centroids using densities from the grid and distance cutoffs for neighbour lists

optional arguments: -h, −−help show this help message and exit
−−author Get information about the authors of this script.
−−version Get version information about this script.
-C [CONFIG], −−config [CONFIG]
Supply an optional Nautilus CONFIG file to control the calculation.
-g [GRIDFORCES], −−gridforces [GRIDFORCES]
Grid.forces file which specifies average parameters of each grid point.
-n [NEIGHCUT], −−neighcut [NEIGHCUT]
The maximum distance (in Angstroms) between grid points to consider them neighbors, recommended value of 1.5
-lt [LOWT], −−lowt [LOWT]
The density threshold to terminate clustering, recommended value of 1.5X greater than bulk
-b, −−benchmark Benchmark the Nautilus subroutines.

An automated protocol, nautilus-protocol:

Runs a default protocol which from the CHARMM DCD trajectory and AMBER topol-

ogy and coordinate file gets clusters of waters in the grid area selected and computes the thermodynamic properties of these clusters. The density clustering defined by the size of the clusters and how much greater than bulk density the cluster is can be controlled within the scripts. Options for the command must be fully specified in the configuration file.

Run the nautilus-protocol command:

usage: nautilus [-h] [-C] [–author] [–version]

Generate clustered centroids of waters and ΔG, ΔH, $-T\Delta S$, and water density

optional arguments:
-h, –help show this help message and exit
-C [CONFIG], −−config [CONFIG]
–author Get information about the authors of this script.
–version Get version information about this script.
-b, –benchmark Benchmark the Nautilus subroutines.

Utilities:

Here a few additional scripts that were implemented are made freely available:

subgrids command used to generate difference grids between two grids. These grids must be of equivalent grid density and size.
avggrids command averages values from 2 or more dx files which again must be of equivalent grid density and size

Summary of the features:
This summarises the current features of the *Nautilus* software. *Nautilus* is available as a plugin for the Sire molecular framework which is now portable over several linux platforms and is provided with unit tests. The software can generate hydration thermodynamics from a DCD trajectory with AMBER topology and coordinate files. It can spatially resolve thermodynamics and can give thermodynamics of any cluster or region desired. DX files can be subtracted to reveal difference isosurfaces which can identify differences between similar ligands and an average command is available for averaging together dx files of replicate runs. It is an outcome of much testing, and debugging over the entirety of the PhD. It will shortly be made available online.